The regulatory capacity of long non-coding RNA in Plasmodium falciparum malaria

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Committee on Higher Degrees in Systems Biology

have examined a dissertation entitled

The regulatory capacity of long non-coding RNA in Plasmodium falciparum malaria

presented by Kate Mariel Broadbent

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: _____

Typed Name: Dr. Andrew Murray

Signature: _____

Typed Name: Dr. Jacquin Niles

Signature: _____

Typed Name: Dr. Matthias Marti

Signature: _____

Typed Name: Dr. Dyann Wirth

Date: June 18, 2014

The regulatory capacity of long non-coding RNA in *Plasmodium falciparum* malaria

A dissertation presented
by

Kate Mariel Broadbent

to
The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in the subject of

Systems Biology

Harvard University
Cambridge, Massachusetts
June 2014

Dissertation Advisors: Prof. Pardis C. Sabeti & Prof. John L. Rinn Kate Mariel Broadbent

The regulatory capacity of long non-coding RNA in *Plasmodium falciparum* malaria

Abstract

The mechanisms underpinning gene regulation in *P. falciparum* malaria remain largely elusive, though mounting evidence suggests a major role for epigenetic feedback. Interestingly, long non-(protein)-coding RNAs (lncRNAs) have been found to play a dominant role in initiating and guiding the transcriptional, epigenetic, and post-transcriptional status of specific loci across a broad range of organisms. LncRNAs are uniquely poised to act co-transcriptionally on neighboring loci, and/or to remain physically tethered at their site of origin, and through sequence-specific binding activities can impart temporal and spatial specificity to ubiquitously expressed nuclear protein complexes. Proteins, on the other hand, must be translated in the cytoplasm, and hence lose memory of their transcriptional origins. Encouraged by these features of lncRNAs, we set out to investigate the regulatory capacity of *P. falciparum* lncRNAs on a genome-wide scale.

First, we surveyed transcriptional activity across approximately one quarter of the *P. falciparum* genome using a custom high-density DNA tiling array. We predicted a set of 60 developmentally regulated intergenic lncRNAs, and found that many of these novel loci

neighbored genes involved in parasite survival or virulence pathways. Remarkably, upon further analysis of intergenic lncRNA properties, we discovered a family of twenty-two telomere-associated lncRNAs encoded in the telomere-associated repetitive element (TARE) region of *P. falciparum* chromosome ends. We found that each lncRNA-TARE was encoded adjacent and divergent to a subtelomeric *var* virulence gene. Moreover, we found that lncRNA-TARE expression was sharply induced between the parasite DNA replication and cell division cycles, that lncRNA-TARE loci contained numerous transcription factor binding sites only otherwise found in subtelomeric *var* promoter regions, and that the GC content and evolutionary sequence conservation of lncRNA-TAREs was similar to that of *P. falciparum* ribosomal RNA.

Next, we set out to assemble *P. falciparum* intergenic lncRNA and antisense RNA transcript structures using state-of-the-art deep sequencing and computational tools. Towards this end, we harvested an unprecedented sample set that finely maps temporal changes across 56 hours of *P. falciparum* blood stage development, and developed and validated strand-specific, non-polyA-selected RNA sequencing methods. This enabled the annotation of over one thousand high-confidence, bona fide lncRNA transcript models, and their comprehensive global analysis. We discovered an enrichment of negatively correlated, tail-to-tail overlapping sense-antisense transcript pairs, suggesting a conserved role for antisense-mediated transcriptional interference in *P. falciparum* gene regulation. We also discovered a highly correlated spliced antisense counterpart to a gene required for sexual commitment, that the expression of an intriguing subset of antisense transcripts significantly dropped during parasite invasion, and that lncRNA-TARE and 'sterile' *var* virulence gene transcription was markedly up-regulated during parasite invasion. Lastly, we predicted over one thousand circular RNAs (circRNAs), and validated six circRNA transcript structures.

Importantly, this thesis work represents the first focused investigation of lncRNAs in *P. falciparum* malaria, with the characterization of a compelling family of telomere-associated lncRNAs and numerous antisense RNAs. The data, methods, and results herein offer exceptional technological advancements coupled with compelling insights into the biology of the devastating human pathogen *P. falciparum* malaria. It is my hope that this work will facilitate future *P. falciparum* lncRNA functional studies and the strand-specific profiling of additional *P. falciparum* samples.

Table of Contents

Acknowledgements

"Here I am, where I ought to be[1]". I especially thank Pardis Sabeti for never fearing that I would get here, and John Rinn for making sure that I did. A more passionate and supportive advisor team does not exist. Thank you both for always keeping it fun, and for believing in me. Thank you Pardis for being my tech guru, for mobile meetings in the Harvard Museum of Natural History or over hot yoga, and for welcoming my family into your lab and home. I think I must hold the world record for number of family members on a lab retreat! And John, without your enthusiasm and vision this project would never have gotten off of its feet. Your critical feedback has been crucial in shaping its course as well.

I am thankful for my Sabeti lab and Rinn lab families. You are all incredible scientists and people, and I have greatly valued your advice and friendship over the years. A big thank you to Loyal Goff and Cole Trapnell for (usually) resisting the urge to run away when I stopped by with a list of computational questions. And to David Hendrickson for always being there to help me out, and for doing so countless times. I thank the Wirth lab members Ulf Ribacke and Daria Van Tyne for being fantastic parasitologists and friends, and for teaching me how to culture parasites. I miss you all already.

I thank my dissertation advisory committee members Dyann Wirth, Jacquin Niles, and Andrew Murray for their guidance throughout my thesis work. I was fortunate enough to work in Jacquin's lab as an undergraduate researcher and to culture parasites in Dyann's lab during my time as a graduate student. I am forever grateful for the many hours Jacquin spent in the lab teaching me, and for encouraging me to pursue graduate school. Almost everything I know about molecular biology, I learned from Jacquin. My time spent rotating in the Wirth lab was my first true contact with parasites, and greatly enriched my understanding of *P. falciparum* biology. It was through this biological lens that I analyzed and interpreted my data.

I am grateful to the National Science Foundation and to the Bill and Melinda Gates Foundation for financially supporting my graduate research. I am also deeply indebted to the Winkler committee for their unexpected and incredible generosity.

Thank you mom and dad for always being there, for working so hard to keep doors open, for choosing horses and travel over a house, for teaching me grit, and, most of all, for loving me, Caroline, Jill, and Tristan so completely. Thank you Jill for your help and hard work in lab. I truly have the best little sister and brother imaginable, and I am so proud of you both. And thank you Indy, Evian, and Sunrise beach for keeping me sane.

And mostly, I want to thank Caroline for bringing more joy and love into my life than I'd ever imagined possible. You have proved to be my biggest inspiration, teacher, and friend yet. Your beautiful mind and heart astound me. Thank you for showing me how to love every day.

[1] Quote by Louise Erdrich in *love life* (love today)

To my sweet Caroline

Chapter 1 | **Introduction**

1. 1 Malaria

There are at least one hundred species of the genus *Plasmodium*, and each parasite species is highly specialized to cause malaria in a specific host: humans, primates, birds, reptiles, or rodents. Transmission of each species depends on a specific vector as well, such as flies or mosquitoes. Of the four *Plasmodium* species that commonly infect humans, *P. falciparum* is by far the most lethal, and is alone responsible for over 90% of malaria-associated deaths. While many drugs can cure *P. falciparum* malaria, and less than 1% of natural infections lead to severe malaria, *P. falciparum* malaria remains one of the most deadly, burdensome, and unchecked infectious diseases of 2014 [1-4]. Annual malaria-associated deaths continue to approach one million [2]. This is largely due to the lack of an effective vaccine, limited access to drugs, the rapid emergence of drug-resistant strains, and, perhaps most significantly, one hundred countries where malaria infection and transmission constantly occur [1,5].

In contrast to cancer, dementia, or other diseases that are often contracted in old age, malaria is disproportionately taxing to the young and the innocent. Children living in sub-Saharan Africa, or other malaria-endemic regions, are particularly vulnerable to severe malaria and malaria-associated death [1,3]. Additionally, pregnancy-associated malaria can be highly deadly to both the mother and developing child [6]. While repeated exposure to *P. falciparum* does not render the parasite incapable of establishing an infection, clinical immunity, or protection from the life-threatening consequences of malaria infection, does typically accumulate with age and exposure to different *P. falciparum* strains [7]. In sum, malaria symptoms may

range in severity from entirely asymptomatic, to headaches, low blood sugar, anemia, and fevers, to kidney failure, severe anemia, respiratory distress, cerebral malaria, and death [3,4].

A single-dose, long-lasting malaria vaccine conferring sterilizing immunity is the ultimate hope for the future [8,9]. However, as neither naturally occurring sterilizing immunity nor elite controllers of *P. falciparum* infection have been documented, this is a daunting challenge. On the other hand, acquired immunity is virtually 100% protective against severe malaria [7], children refractory to severe malaria have recently been reported [10], and pregnancy-associated infections are known to involve a moderately conserved parasite ligand [11]. Thus, development of a vaccine conferring protective, clinical immunity to malaria-naïve individuals or during pregnancy may be more immediately feasible [12]. Transmission-blocking vaccine (TBV) development is another intriguing consideration, as are novel approaches that target the vector stages and/or vectors specifically [13,14]. To develop and deploy anti-malarial strategies faster than the parasite and vector thwart them, and potentially faster than nature has provided them, will require a sustained commitment to basic, field, and translational malaria research, cross-disciplinary approaches, improvements in local health care systems and surveillance, and cross-talk between the various groups [5,15].

1.2 The *P. falciparum* life cycle

The *P. falciparum* life cycle is complex, and involves multiple stage transformations in both a human host and an anopheline mosquito vector **[Figure 1.1]** [16-18]. During a blood meal, only a few infectious parasite forms may be transferred from mosquito to human and/or vice-versa [19]. However, the parasite utilizes a unique form of schizogony involving multiple rounds of asexual replication to rapidly expand population numbers after both bottlenecks [20].

In the human, an asexual liver schizogony stage precedes the asexual blood schizogony stage. The latter blood stage occurs in host red blood cells (RBCs), repeats every 48 hours, and causes all clinical malaria symptoms [3,4]. In the mosquito, when both a male and a female gametoycte are ingested, rapid sexual replication events are followed by asexual schizogony [20]. The *P. falciparum* parasite thus also uses sexual replication in the mosquito vector to recombine genetic information and to accumulate adaptations. Importantly, while sexual replication occurs in the mosquito, the prerequisite switch from asexual to sexual stage gametocyte development, or gamete precursor development, takes place within the human host [21].

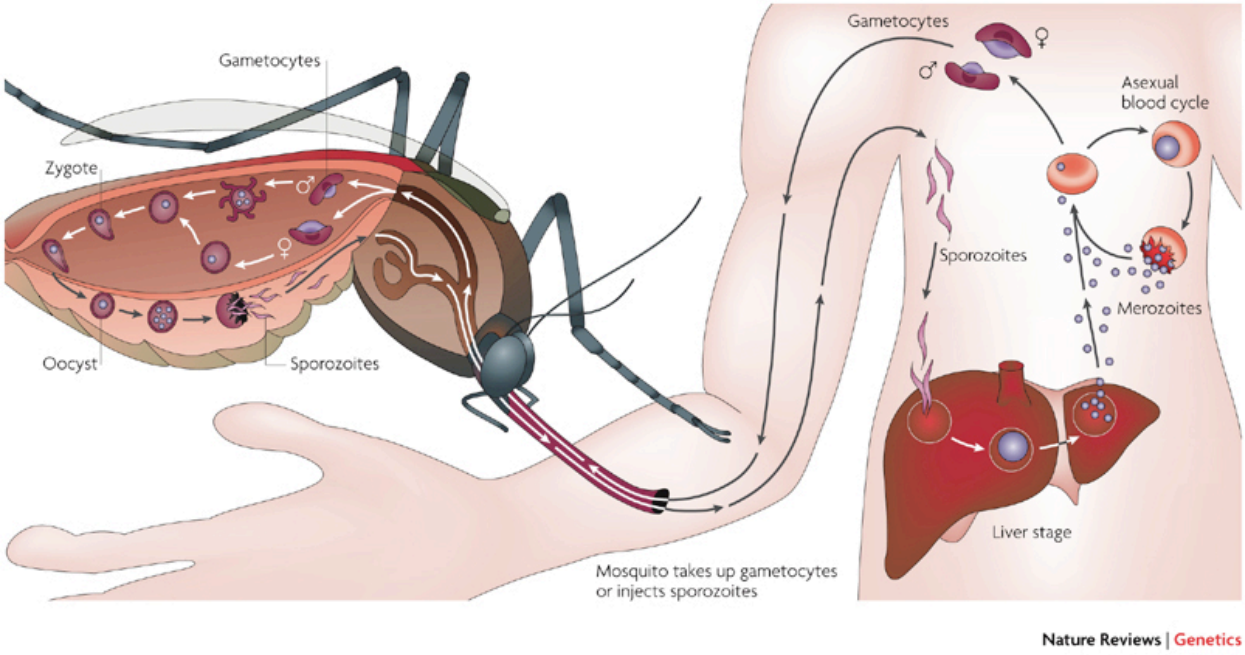


Figure 1.1 | Overview of the *P. falciparum* malaria life cycle and host transfer. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [16], copyright 2007.

Traversing the parasite cycle in more detail reveals the remarkable fine-tuning of developmental programs and parasite processes necessary for parasite survival. From the perspective of an infectious mosquito bite, the first stage in the parasite life cycle is the rapid traversal of approximately 15-40 haploid sporozoites from below the skin layer into the human

blood stream [22]. Within thirty minutes, a handful of these sporozoites successfully invade liver cells [23]. Over the course of two weeks, each sporozoite then duplicates its genome and completes approximately 13 to 14 rounds of alternating DNA synthesis and mitosis (S/M) [20]. Subsequently, thousands of haploid daughter merozoites are released into the blood stream, streamlined to find, invade, and feed on RBCs [24]. Within minutes, these liver-stage merozoites accomplish their mission, marking the end of the clinically asymptomatic pre-erythrocytic phase.

During the next phase of parasite development, known as the clinically symptomatic intraerythrocytic developmental cycle (IDC), the parasite repeatedly progresses through a ring (0-24 hours post-invasion [hpi]), a trophozoite (24-36 hpi), and a schizont (36-48 hpi) stage in host RBCs [25]. During the ring stage, the parasite flattens into a thin ring-like form, and begins digesting RBC cytosol and catabolizing hemoglobin [26]. Notably, the parasite also exports hundreds of proteins to repurpose and remodel the normal RBC machinery and RBC surface, such that the later trophozoite and schizont stages are able to adhere to microvasculature, sequester in various host tissues, and avoid splenic clearance [25,27]. During the trophozoite stage, the parasite becomes notably rounder and larger; DNA replication begins, and metabolic activity and protein synthesis dramatically increase [26]. During the schizont stage, approximately 3 to 4 rounds of S/M occur, followed by nuclear division, packaging, and the release of up to 32 haploid daughter merozoites into the blood stream [20]. The vast majority of these blood-stage merozoites are committed to initiate a new IDC [28,29].

Interestingly, however, a poorly defined switch in developmental program occurs in a small subset of asexually cycling blood stage parasites, such that all of the daughter merozoites from a given schizont become sexual gametocytes. Moreover, all of the daughter merozoites become either male or female gametocytes [21,28,30]. Sex ratios can vary considerably, though

a three-fold higher female to male gametocyte ratio is typically observed *in vivo* [31]. Mature gametocyte development takes approximately 8-12 days in *P. falciparum*, which is significantly longer than gametocytogenesis in any other *Plasmodium* species (except *P. reichenowi*), and progresses through five distinct morpholigical stages [32].

Stage I gametocytes (up to 40 hpi) are hard to discern from asexual parasite stages by eye, though an antibody to the membrane protein *Pfs16* or the cytoplasmic RNA-binding protein *Pfg27/25* can be used to identify early gametocytes [33,34]. During this stage, the parasite exports hundreds of proteins, again repurposing and remodeling the normal RBC machinery and surface [35,36]. These changes confer novel adhesive properties to the host RBC, and allow gametocytes to strategically sequester and avoid splenic clearance until they reach maturity [37,38]. Stage II-V gametocytes can be distinguished by eye. The growing gametocyte lengthens to occupy the majority of the RBC, and from stage IV onwards, male and female gametocytes are visibly distinct [21]. Mature stage V gametocytes are often crescent shaped, and circulate in the peripheral blood stream, waiting to be ingested by a mosquito [39].

When a male gametocyte is ingested, within fifteen minutes it completes three rounds of S/M, nuclear division, and cytokinesis to create eight male gametes [20,37]. When a female gametocyte is ingested, it also rapidly develops into a female gamete. The fusion between a male and a female gamete then yields a diploid zygote. The zygote soon develops into a motile ookinete, which must traverse the mosquito midgut epithelial wall, and embed itself into the basal lamina as an oocyst [18]. Over the next few days, the oocyst undergoes 10 to 11 rounds of S/M, which is followed by a mass nuclear division and the release of thousands of haploid daughter sporozoites [20]. These sporozoites then migrate to the mosquito salivary gland, ready

to be transmitted to a new host. All together, the mosquito phase of the *P. falciparum* life cycle takes approximately 10-14 days [19].

1. 3 Pathogenesis of *P. falciparum* infection

There are three major severe malaria syndromes: severe anemia, cerebral malaria, and respiratory distress [3,4,40]. Respiratory distress presents in 40-60% of severe malaria cases, and is caused by reduced oxygen delivery to tissues and metabolic acidosis [4]. As only about 1% of malaria infections progress to severe malaria, severity of disease and clinical outcome is likely caused by a complex combination of parasite, host, geographic, and social factors [3]. In terms of *P. falciparum* factors, different parasite strains have different multiplication rates, drug sensitivities, and repertoires of antigen and invasion ligands [41-44]. Protective host factors include previous exposure, lack of pregnancy, otherwise good health, and certain genetic backgrounds [3]. For example, sickle cell trait *or* α-thalassaemia alleles greatly reduce the incidence of severe malaria [45-47]. Access to treatment, political stability, transmission intensity, and other geographic, cultural, and socioeconomic factors are also important variables in severe malaria progression [3].

Interestingly, *P. falciparum* is responsible for the vast majority of severe malaria cases and deaths. This heightened pathogenicity is largely due to the redundancy and diversity of invasion and adhesion ligands that have expanded in the *P. falciparum* genome, as well as their precise regulation [3,4,48-50]. For example, in contrast to the human malaria parasite *P. vivax*, which historically can only invade reticulocytes with the Duffy blood group (Fy) antigen [51], *P. falciparum* can invade most types of RBCs [24,52]. It can also use sialic acid dependent or independent invasion pathways [52]. The multi-gene *P. falciparum Erythrocyte Binding Antigen*

(*PfEBA*) and *P. falciparum Reticulocyte Binding Antigen Protein homolog* (*PfRh*) families encode sialic acid dependent and independent invasion ligands, respectively, with many redundant family members [53-55].

The host splenic clearance system removes RBCs with altered deformability, which in itself is an effective way to clear parasite-infected RBCs from the blood stream [56]. However, *P. falciparum* encodes a hyper-variable family of sixty *P. falciparum* Erythrocyte Membrane Protein 1 (PfEMP1)-encoding *var* genes [57]. This family contributes a key RBC surface adhesin during the asexual blood stage, and gives *P. falciparum* unique cytoadherence properties as compared to other human *Plasmodium* species. PfEMP1 adhesins can bind to protein and carbohydrate receptors on the surface of endothelial cells, such as CD36, ICAM1, PECAM, CR1, heparin sulfate (HS), and placental chondroitin sulfate A (CSA) [4,49]. The ability of *P. falciparum* to hide out and sequester in various host organs, such as the heart, lungs, kidneys, liver, brain, and/or placenta, is critical for *P. falciparum* pathogenesis and persistence [3,4,48-50]. Developing sexual stages are also known to cytoadhere until reaching maturity, although the adhesion molecules involved are less clear [37,38].

1.4 Gene regulation in *P. falciparum*

Approximately 5,300 protein-coding genes are encoded in the 24 megabase *P. falciparum* genome, in addition to clusters of non-(protein)-coding ribosomal RNA (rRNA), transfer RNA (tRNA), and small nucleolar RNA (snoRNA) genes [58]. *P. falciparum* genes are distributed onto fourteen linear chromosomes, and two small circular apicoplast and mitochondrial chromosomes [58]. Remarkably, approximately 80% of the *P. falciparum* genome is periodically regulated during the asexual blood stage [59,60]. By comparison, only 15% of yeast and human

genes are periodically regulated during the cell cycle [61,62]. Moreover, gene ontology analysis of co-expressed *P. falciparum* genes has suggested that the periodic blood stage cascade dictates a 'just in time' manufacturing process, or in other words that the steady-state mRNA levels of most genes strongly correlates with the timing of their function [59,60].

Notably lacking from the *P. falciparum* genome are the RNA interference machinery components and microRNAs, as well as many classes of sequence-specific transcription factors [63,64]. However, the parasite does have 27 plant-like ApiAP2 DNA-binding factors with putative sequence-specific transcription factor activities, an intriguing abundance of CCCH-type zinc finger proteins, and several classes of RNA-binding proteins [64-67]. CCCH-type zinc finger proteins are known to regulate mRNA stability and translation in other eukaryotes, though their roles in *P. falciparum* have not yet been functionally investigated [68]. To date, numerous ApiAP2 DNA-binding factors have been experimentally knocked out or knocked down, and have been shown to play an important role in *P. berghei* ookinete formation, oocyst to sporozoite maturation, liver stage development, and gametocyte development [69-72]. ApiAP2 factors have been implicated in *P. falciparum* gametocyte development and virulence gene regulation as well [73-75].

The *P. falciparum* genome encodes a wide array of conserved chromatin reader, writer, and eraser enzymes, and both conserved and unusual histone variants [76]. Additionally, at least 44 different histone tail post-translational modifications (PTMs) have been observed using mass spectrometry [77]. Genome-wide chromatin immunoprecipitation ChIP-chip and ChIP-seq studies have shown that the distribution of many PTMs is dynamic [78-80]. Specifically, the parasite has at least four histone acetyltransferases (HATs), at least ten histone lysine methyltransferases (HKMTs), and at least five histone deacetylases (HDACs) [76]. Common

histone tail PTMs include acetylation, methylation, phosphorylation, ubiquitination, poly-ADP-ribosylation, and sumoylation [77]. Complex combinatorial effects of histone variants and PTMs are hypothesized [76]. Notably, more activating PTMs have been observed in *P. falciparum* than silencing PTMs, which is consistent with reports showing that approximately 90% of the *P. falciparum* genome is in a transcriptionally permissive, euchromatic state [77,80]. The other 10% of the genome is marked by H3K9me3 and *P. falciparum Heterochromatin Protein 1* (*PfHP1*) [80-82], and is thought to represent heterochromatic islands that help to maintain multi-gene family members in a silenced state [83]. Indeed, the vast majority of *PfHP1* occupancy overlaps with multi-gene families [81].

Taken together, it is likely that both transcription factors and chromatin state are largely important in the precise control of *P. falciparum* network-wide and local gene expression programs [76,84-86]. It is also becoming increasingly clear that post-transcriptional modifications, such as mRNA stability and translational repression, are global themes in *P. falciparum* gene regulation [17,87-90]. Early case studies pointed to a role for post-transcriptional regulation in sexual differentiation events [91,92] and mitochondrial RNA processing [93]. Shock et al. then investigated mRNA decay rates during the asexual blood stage, and found a profound global extension in the rate of mRNA decay during the blood stage [88]. Other groups have compared steady-state mRNA levels to protein levels or ribosome-associated RNA levels during the blood stage, and have reported that approximately 30% of *P. falciparum* genes are subject to significant translational repression [60,87].

Focusing in on sexual stage gametocyte development, translational repression is likely to be key to the rapid developmental events triggered in the mosquito midgut [94]. Experiments in *P. falciparum* suggest that there is no DNA replication during gametocyte development [95], and

that RNA synthesis significantly declines after day six [96,97]. However, from studies involving the rodent malaria parasite *P. berghei*, it appears that the female gametocyte stockpiles mRNAs during the later gametocyte stages in a ribonucleoprotein complex containing the DDX6-class RNA helicase DOZI (development of zygote inhibited) and several additional factors [94,98]. Potentially hundreds of mRNAs, including *Pfg25* and *Pfg28*, are then rapidly translated in the mosquito in response to environmental stimuli [60,94,99]. Knocking out the Puf-family RNA-binding protein (*PfPuf2*) has also been shown to disproportionately repress male gametocyte development [100]. Puf-family proteins are known to regulate translation and RNA stability in other eukaryotes [101].

Additionally, case studies point to the important roles of PTMs in regulating *P. falciparum* protein activity. For example, proteolytic cleavage of two *P. falciparum* phosphatases (PP7 and PP2b) confers unique functional properties to the processed enzymes, as does cleavage of an ApiAP2 DNA-binding factor during the schizont stage [75]. As another example, at least five enolase post-translationally modified isoforms have been described, which seem to preferentially localize to different subcellular compartments [102]. Out of 54 schizont stage proteins studied by mass spectrometry, at least half were observed to have multiple isoforms [90]. This set included enolase again, as well as actin I, eukaryotic initiation factors eIF4A and eIF5A, and numerous heat shock proteins. Importantly, Zhang et al. have recently shown that eIF2α phosphorylation by the *PfPK4* kinase is essential for blood stage development in *P. berghei*, and have hypothesized that this interaction is likely to be broadly conserved amongst *Plasmodium* species [103].

1.5 Signatures of multi-gene family regulation

The expression pattern of multi-gene families such as the PfEMP1-encoding *vars*, *PfEBAs*, and *PfRhs* has been well studied during *P. falciparum* blood stage development [104]. In the case of *PfEMP1*, expression is mutually exclusive, meaning that only a single PfEMP1 variant is displayed on the RBC surface at a given time [105,106]. PfEMP1 expression choice is also largely clonally inherited during the blood stage by daughter parasites [104-106]. However, *in vivo* experiments have shown that up to 18% of daughter parasites may switch to express a different PfEMP1 variant during the initial stage of infection [107]. A similar program of mutually exclusive, heritable gene expression has been documented for the two neighboring *Cytoadherence Linked Asexual Gene 3* (*CLAG3*) loci [108,109]. The *CLAG3* loci encode nutrient uptake channel proteins [110,111]. Expression of *PfEBA* and *PfRh* genes is clonally variant and heritable as well, though a few genes from each family are typically co-expressed at any given time [104]. Whether the basis for locus-specific expression and expression switching of multi-gene families is purely stochastic, or if it involves host-parasite interactions is debated [105]. Regardless, it is clear that this strategy is key to *P. falciparum* pathogenicity and immune evasion [3,24,49,50].

The PfEMP1-encoding *var* genes have been studied in the most detail out of *P. falciparum* multi-gene families, and their regulation involves many layers: epigenetic memory passed from generation to generation, conserved sequence elements encoded in *var* gene promoters and introns, and subnuclear localization events, amongst others (reviewed in [105]). Specific histone modifications are known to mark active, poised, and silenced *var* loci [82,112]. Additionally, Volz et al. have implicated *PfSET10*, one of many *P. falciparum* HKMTs, in

maintaining the H3K4me3 histone modification over the poised *var* gene [113]. Deitsch et al. and others have shown that strict promoter-intron pairing is required for mutually exclusive *var* gene regulation, and that distinct nuclear protein complexes bind pairing elements (PE) located in these regulatory regions [114-117]. Regarding subnuclear localization, Ralph et al. and others have used fluorescence in situ hybridization (FISH) to show that both subtelomeric and internal *var* genes cluster at the nuclear periphery in 4-7 distinct foci, and that *var* gene activation involves subnuclear repositioning events [118,119].

Providing a molecular basis that connects many of these results, Zhang et al. have recently characterized a nuclear actin-protein complex that recognizes a short repeat region present in internal *var* gene introns [74]. This 18 base pair element is termed iNPE (intron nuclear protein-binding element), and is directly bound by the ApiAP2 DNA-binding factor *Pf3D7_1107800*. Specifically, this group showed that iNPE is sufficient to recruit episomes to the nuclear periphery, and that actin polymerization is essential for spatial repositioning and *var* gene regulation [74]. Another recent study by Coleman et al. has shed new light on variant gene expression in *P. falciparum* as well [53]. This group specifically investigated the timing of *PfRh4* nuclear repositioning and chromatin accessibility, and found that two distinct regulatory steps are involved in activation of this locus. Notably, Coleman et al. found that nuclear repositioning of *PfRh4* to a transcription-permissible site is necessary but not sufficient for transcriptional activation of this locus. In other words, nuclear repositioning does not in itself cause transcriptional activation or promoter accessibility, but it does markedly increase the frequency of transcriptional activation events. This two-step activation model may also hold for *var* gene regulation [53].

1.6 Outstanding gene regulatory questions

One important question is the extent to which the *P. falciparum* blood stage transcriptome is hard-wired versus reactive to environmental factors, and how putative response signals may be integrated. In support of the notion that the *P. falciparum* transcriptome is rigid, the blood stage expression cascade of geographically diverse *P. falciparum* isolates is remarkably similar [120,121]. This is true for different *Plasmodium* species as well [121]. Moreover, lethal dosage with chloroquine or the antifolate drug WR99210 yielded a minimal transcriptome-wide effect [122,123]. Arguing for transcriptome plasticity, Chaal et al. showed that treatment with the HDAC inhibitor apicidin affected the stage-specific expression of approximately half of the *P. falciparum* transcriptome [124]. Other groups have shown that treatment with artesunate, an artemesinin derivative, or the bisthiazolium compound T4 elicited major transcriptional effects as well [125,126]. Notably, *P. falciparum* responded in an expected manner to febrile temperatures or hyperoxia exposure, up-regulating protective antioxidant systems and down-regulating glycolysis in the latter case [127,128]. As speculated by Kato et al., it is likely that *P. falciparum* has lost some of the response mechanisms of free-living eukaryotes, but retained the ability to respond to key growth and environmental perturbations. Understanding the parasite elements that guide such responses may lead to novel blood stage therapeutic targets [129].

How gametocyte commitment is regulated remains an open question, and again the degree to which environmental triggers are involved is debated [37]. A growing list of studies, primarily performed *in vitro*, have shown that environmental factors such as RBC age, hypoxia, drug exposure, and overcrowding can impact commitment rates [13,37,130]. Most recently, roles

for host microRNAs and/or microvesicles, which bleb from infected RBCs prior to parasite egress, have been connected to *in vitro* gametocyte commitment rates as well [131-133]. In terms of molecular insights into gametocyte commitment, two nuclear proteins have been strongly implicated: *P. falciparum* Gametocyte Developmental Protein 1 (PfGDV1) and an ApiAP2 DNA-binding factor [73,134]. The *PfGDV1* locus was discovered due to reports showing that parasites cultured *in vitro* often spontaneously deleted this genomic region, overtook cultures, and could no longer produce visible gametocytes [135-137]. As each committed gametocyte represents a dead end *in vitro*, this growth advantage suggested that gametocyte commitment was likely blocked in these deletion strains. Recent random mutagenesis and complementation experiments performed by Ikadai et al. have also linked at least four additional genes to gametocyte commitment and/or early developmental events [138]. Notably, while Kafsack et al. have proposed that stochastic activation of the ApiAP2 factor *Pf3D7_1222600* controls baseline gametocyte commitment [73], this model is likely incomplete. Knockdown of *Pf3D7_1222600* was not found to alter the expression of *PfGDV1* or the other gametocyte-implicated genes reported by Ikadai et al. [73,134,138], suggesting additional complexity.

Albeit a rapidly advancing field, mutually exclusive *var* gene regulation is far from understood. For example, while maintenance of heterochromatin marks and clustering at the nuclear periphery appear necessary for *var* gene silencing, how *var* genes reposition to a transmission-permissible site remains unclear[105]. Moreover, observations implicating non-coding RNA and RNA-binding proteins in *var* gene regulation have not been experimentally investigated [139]. In 2003, Kyes et al. described ‘sterile’ non-coding RNAs transcribed from bidirectional *var* intron promoters [140]. Epp et al. then showed that these non-coding transcripts associated with chromatin [141]. Recently, RNA-binding proteins have been identified

associated with both *PfSET10* and the iNPE using pull-down experiments [74,113]. In fact, ten out of the thirteen top candidate proteins in complex with the iNPE have predicted RNA-binding domains [74].

1.7 The regulatory capacity of lncRNA

Long non-(protein)-coding RNAs (lncRNAs) represent a diverse class of regulatory molecules that have been found to modulate gene expression from every conceivable angle, often in a uniquely site-specific manner [142]. To date, detailed experimental dissection of lncRNAs and their target genes has been performed for numerous lncRNAs **[Table 1.1]** (Reviewed in[143-145]). However, the mammalian X-inactive-specific transcript (Xist), which represents one of the first bona fide lncRNAs to be discovered, continues to be one of the most experimentally investigated lncRNAs [146-150]. Numerous studies have demonstrated the essential role for Xist, a 17-20 kb lncRNA expressed from the inactive X chromosome, in X chromosome inactivation (XCI) [149,151]. Moreover, many additional lncRNAs in the X-inactivation center (*Xic*) regulate XIC through a series of RNA-based switches, and together contribute to the pairing of *Xic* on different chromosomes, the recruitment of numerous chromatin modifying enzymes, and the spread and maintenance of heterochromatic silencing marks along the entire inactive X chromosome **[Figure 1.2]** [152-156]. In brief, Xist is positively regulated by the upstream Jpx lncRNA and negatively regulated by its antisense transcript Tsix [157-159]. X-inactivation intergenic transcription elements (Xite) positively regulate Tsix [160].

Table 1.1 | Examples of lncRNA-mediated gene expression control. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology [143], copyright 2013.

lncRNA	Function	Mechanism
Regulation of mRNA transcription		
XIST	X inactivation	Chromatin-mediated repression
HOTAIR	Repression at the *HOXD* locus	Chromatin-mediated repression
HOTTIP	Activation at the *HOXA* locus	Chromatin-mediated activation
KCNQ1OT1	Imprinting at the *KCNQ1* cluster	Chromatin-mediated repression
ANRIL	Repression at the *INK4b–ARF–INK4a* locus	Chromatin-mediated repression
AIRN	Imprinting at the *IGF2R* cluster	Chromatin-mediated repression, transcription interference
IME4 antisense	Repression of *IME4* mRNA	Transcription interference
IRT1	Repression of *IME1* mRNA	Chromatin-mediated repression
GAL10 lncRNA	Repression of *GAL1* and *GAL10* mRNAs	Chromatin-mediated repression
PHO84 antisense	Repression of *PHO84* mRNA	Chromatin-mediated repression
ICR1	Repression of *FLO11* mRNA	Modulation of transcription factor recruitment
PWR1	Activation of *FLO11* mRNA	Modulation of transcription factor recruitment
SRG1	Repression of *SER3* mRNA	Nucleosome remodelling
fbp1 ncRNA	Activation of *fbp1*	Chromatin remodelling
LINOCR	Activation of lysozyme mRNA	Nucleosome remodelling
Alu repeat-containing RNA	Transcriptional repression during heat shock	Inhibition of Pol II
HSR1	Activation of the HSF1 transcription factor	Allosteric activation together with eEF1A
Non-coding *DHFR*	Transcriptional repression of *DHFR*	Inhibition of pre-initiation complex formation
GAS5	Repression of glucocorticoid receptor-mediated transcription	DNA mimicry
EVF2	Transcriptional activation of DLX2 targets, transcriptional repression of MeCP2 targets	Recruitment of DLX2 or MeCP2
CCND1 promoter RNA	Repression of *CCND1* transcription	Allosteric activation of TLS
NRON	Repression of NFAT-mediated transcription	Inhibition of transcription factor nucleocytoplasmic shuttling

Table 1.1 (Continued) | Examples of lncRNA-mediated gene expression control.

Regulation of mRNA processing		
Neuroblastoma *MYC* (NAT)	Inhibition of neuroblastoma *MYC* intron 1 splicing	Unknown mechanism involving the inhibition of splicing via RNA–RNA duplex formation
*Rev-ErbA*alpha	Inhibition of the *c-ErbAalpha 2* splice isoform	Unknown mechanism involving the inhibition of splicing via RNA–RNA duplex formation
ZEB2 (NAT)	Activation of *ZEB2* translation	Unknown mechanism involving regulated splicing of an IRES-containing intron
MALAT1	Ser/Arg splicing factor regulation	Scaffolding of subnuclear domains
Sas10 mRNA 3′ UTR	Repression of *Rnp4F* mRNA	Unknown mechanism involving RNA editing
Modulation of mRNA post-transcriptional regulatory pathways		
Antisense *UCHL1*	Upregulation of UCHL1 protein production	SINE2B element-mediated translational upregulation
KCS1 antisense	Production of truncated KCS1 protein	Unknown mechanism involving base pairing
1/2-sbsRNA 1	Down-regulation of *SERPINE1* and *FLJ21870* mRNAs	Staufen-mediated decay through Alu element base pairing
BACE1AS	Up-regulation of *BACE1*	Stabilization of *BACE1* mRNA by blocking miRNA-induced repression
LINCMD1	Control of muscle differentiation through upregulation of MAML1 and MEF2C transcription factors	Sequestration of miRNAs
HULC	Downregulation of miRNA-mediated repression	Sequestration of miRNAs
PTENP1 pseudogene	Upregulation of PTEN	Sequestration of miRNAs
IPS1	Downregulation of miRNA-mediated repression	Sequestration of miRNAs
CDR1as	Downregulation of miRNA-mediated repression	Sequestration of miRNAs

1/2-sbsRNA1, half-STAU1-binding site RNA 1; *AIRN*, antisense of *IGFR2* non-coding RNA; *BACE1AS*, beta-site APP-cleaving enzyme 1 antisense; *CCND1*, cyclin D1; *CDR1as*, CDR1 antisense; *DHFR*, dihydrofolate reductase; *fbp1*, fructose-1,6-bisphosphatase 1; eEF1A, eukaryotic elongation factor 1A; *FLO11*; *GAS5*, growth arrest specific 5; *HOTAIR*, HOX transcript antisense RNA; *HOTTIP*, *HOXA* transcript at the distal tip; *HOX*, homeobox cluster; HSF1, heat shock factor 1; *HSR1*, heat shock RNA 1; *HULC*, highly upregulated in liver cancer; *IGF2R*, insulin-like growth factor 2 receptor; *IME*, inducer of meiosis; *IPS1*, INDUCED BY PHOSPHATE STARVATION 1; *IRES*, internal ribosome entry site; *IRT1*, *IME1* regulatory transcript 1; *KCNQ1*, potassium voltage-gated channel, KQT-like subfamily, member 1; *KCNQ1OT1*, KCNQ1 opposite strand or antisense transcript 1; *LINOCR*, *LPS*-inducible non-coding RNA; lncRNA, long non-coding RNA; *MALAT1*, metastasis associated lung adenocarcinoma transcript 1; MAML1, mastermind-like 1; MeCP2, methyl CpG binding-protein 2; MEF2C, myocyte enhancer factor 2C; miRNA, microRNA; NAT, natural antisense transcript; ncRNA, non-coding RNA; NFAT, nuclear factor of activated T cells; *NRON*, non-coding repressor of NFAT; Pol II, RNA polymerase II; *PTENP1*, phosphatase and tensin homologue; *Rnp4F*, RNA-binding protein 4F; TLS, translocated in liposarcoma; *UCHL1*, ubiquitin carboxyl-terminal esterase L1; UTR, untranslated region; *XIST*, X inactivation-specific transcript; *ZEB2*, zinc-finger E-box binding homeobox 2.

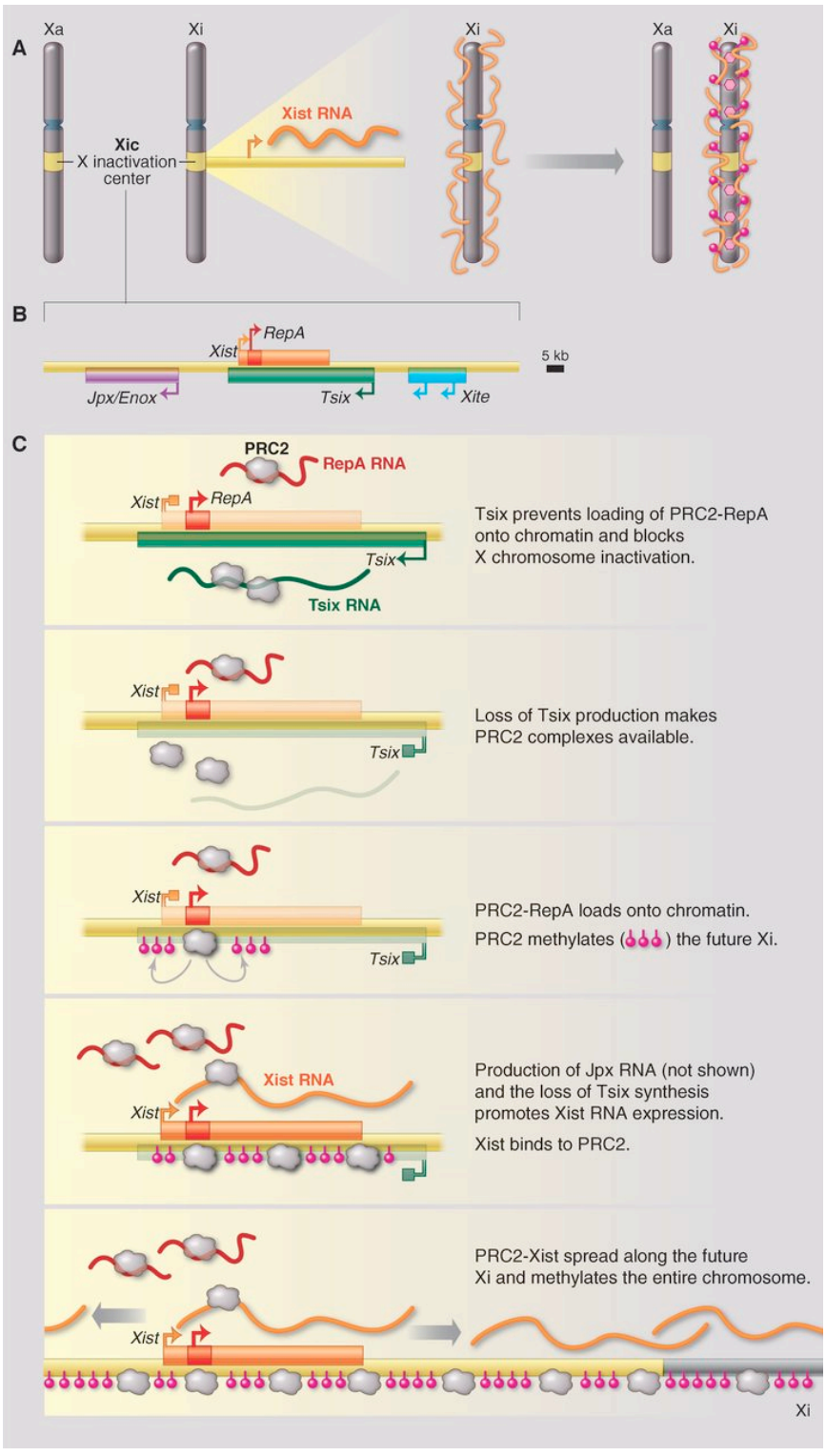


Figure 1. 2 | Non-coding genes at the *Xic*. Reprinted from Science, Volume 338, Page 1436 [142], with permission from the American Association for the Advancement of Science. *Epigenetic Regulation by Long Noncoding RNA.* Jeannie T. Lee. Copyright 2012.

To put the XCI cascade into the larger picture of mammalian development, Oct4 binding activates Tsix and Xite expression [161,162]. The pluripotency-associated transcription factor Oct4 is thus able to initiate local effects at the *Xic* through Oct4-responsive lncRNAs. Additionally, Oct 4 is able to initiate genome-wide transcriptional programs in response to bone morphogenetic protein signaling or leukemia inhibitory factor [142,161,162]. Intriguingly, it has also recently been suggested that Oct4 is itself transcriptionally regulated by the lncRNA asOct4-ps5, an antisense transcript to Oct4-pseudogene 5 [163]. Taken together, XCI demonstrates the complex collaboration between site-specific lncRNAs, transcription factors, and chromatin modifiers, which together provide spatial and temporal specificity during mammalian development. The lncRNAs in the *Xic* are amongst the most well studied lncRNAs, and continue to provide insight into the mechanisms by which lncRNAs may exert regulatory control [150,152,153].

Additional well-characterized mammalian lncRNAs include lncRNAs involved in the epigenetic phenomenon known as imprinting. The silenced alleles in imprinted gene clusters are marked by DNA methylation and repressive histone modifications, reminiscent of the inactive X chromosome [164,165]. However, in this case, monoallelic gene silencing is dependent on parental origin. For example, the insulin-like growth factor type-2 receptor (*Igf2r*) lies within an imprinted gene cluster. Only the maternal *Igf2r* allele is expressed [166]. Transcription of Antisense to Igf2r RNA non-coding (*Airn*) from the second intron of paternal *Igf2r* alleles induces the silencing of paternal *Igf2r* and two neighboring alleles [167]. As another example, *Kcnq1* lies within a one megabase imprinted gene cluster [168]. The antisense *Kcnq1ot1* RNA from intron 10 of paternal *Kcnq1* alleles induces the silencing of paternal *Kcnq1* and neighboring alleles through interactions with a histone methyltransferase (HMT) and the polycomb repressive

complex 2 (PRC2) [169]. LncRNAs have also been implicated in Alzheimer's disease [170], Parkinson's disease [171], Fragile X syndrome [172], and various forms of cancer [173,174].

1.8 LncRNA case studies in model organisms

The model budding yeast *Saccharomyces cerevisiae* is a unicellular eukaryote of comparable genome size and gene density to *P. falciparum* [58]. Interestingly, *S. cerevisiae* also lacks the RNA interference pathway [175,176]. It is thus possible that the functional roles of lncRNAs in *S. cerevisiae* may be particularly predictive of lncRNA functions in *P. falciparum* [177]. The list of functionally validated lncRNAs in *S. cerevisiae* continues to grow, with regulatory roles ranging from mating-type control of meiosis entry to blocking serine biosynthesis [177-180].

In *S. cerevisiae,* entry into meiosis requires the function of many genes and the convergence of many external signals [181]. Nutritional and respiratory signals include starvation, the lack of fermentable sugar and nitrogen sources, and the presence of non-fermentable carbon sources. Furthermore, only diploid mating type locus MATa/α cells can undergo meiosis and sporulation. The a1/α2 heterodimer, unique to diploid cells with both a MATa and MATα allele, represses haploid functions. The meiotic cascade is set into motion by a family of *IME* (inducer of meiosis) genes. The promoter of *IME1* is highly regulated, and integrates both environmental and mating-type control signals. Another gene required for meiotic entry is *IME4*, a putative RNA methyltransferase [181].

Interestingly, lncRNAs regulate both *IME1* and *IME4* in *S. cerevisiae* **[Figure 1.3]** [179-181]. In haploid MATa and MATα cells, *IME4* antisense RNA is constitutively transcribed and *IME4* sense RNA is silenced [179]. In diploid MATa/α cells, the inverse relationship is true.

Through a series of genetic perturbation assays, Hongay et al. discovered that *IME4* sense and antisense transcription inhibit each other in *cis*, and that the relative strength of their promoters is cell-type specific. Namely, while the *IME4* antisense promoter is stronger in haploid cells, binding of the 1a/2α heterodimer (present only in MATa/α diploid cells) at a conserved site downstream of *IME4* favors the *IME4* sense promoter. Taken together, this system ensures that only diploid cells are competent for meiosis entry [179]. At the *IME1* locus, the repressor of meiosis 1 protein (Rme1p) binds to the *IME1* promoter in haploid cells [181]. Rme1p binding activates transcription of IME1 regulatory transcript 1 (*IRT1*), which spans the *IME1* promoter [180]. IRT1 transcription recruits HMT (Set2) and HDAC (Set3C) complexes to the *IME1* promoter, which in turn establishes a repressive chromatin state that prevents further binding by other transcription factors [180].

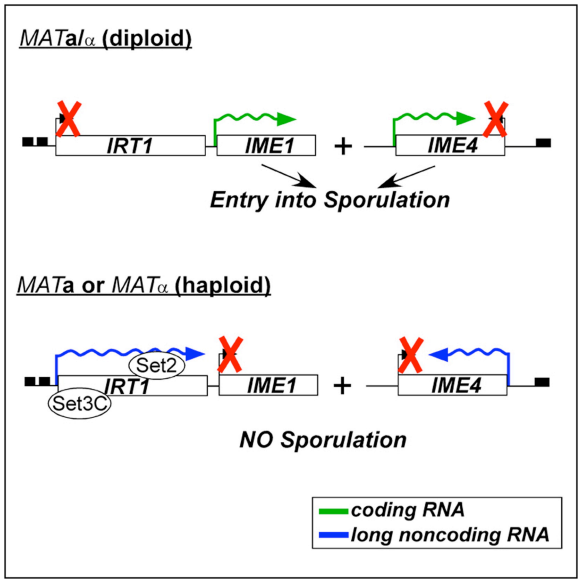


Figure 1. 3 | Mating-type control of meiosis entry is regulated by two lncRNAs. Reprinted from Cell, Volume 150, Issue 6, with permission from Elsevier. *Transcription of Two Long Noncoding RNAs Mediates Mating-Type Control of Gametogenesis in Budding Yeast*. Folkert J. van Werven, Gregor Neuert, Natalie Hendrick, Aurélie Lardenois, Stephen Buratowski, Alexander van Oudenaarden, Michael Primig, Angelika Amon. Copyright 2012.

At the *S. cerevisiae SER3* locus, which encodes an enzyme involved in serine biosynthesis, transcription of SER3 regulatory gene 1 (*SRG1*) represses *SER3* transcription [178]. The *SRG1* locus overlaps with the *SER3* upstream regulatory sequence, and functions by interfering with activator binding at the *SER3* promoter. SRG1 is only transcribed when cells are grown in the presence of serine, however, such that *SER3* expression may be induced when serine biosynthesis is necessary [178]. LncRNAs also regulate the expression of enzymes involved in galactose biosynthesis (GAL10 and GAL4) [182], the adhesion-conferring FLO11 cell wall glycoprotein [183,184], and the inorganic phosphate transporter PHO84 [185].

Fundamental roles for lncRNAs have been documented in numerous other model organisms as well [186,187]. In the plant *Arabidopsis thaliana*, for example, exposure to winter cold triggers the epigenetic silencing of the flowering locus c (*FLC*) gene, and gives *A. thaliana* the ability to flower in the spring [188]. This process is called vernalization. The COLD ASSISTED INTRONIC NON-CODING RNA (COLDAIR) is transcribed from the first *FLC* intron after prolonged cold exposure, and is required for long-term maintenance of *FLC* repression and vernilization [189]. COLDAIR associates with the HMT component of PRC2, and likely targets PRC2 to the *FLC* locus [189]. In the fruit fly *Drosophila melanogaster*, sex chromosome dosage compensation is achieved by up-regulating the single male X chromosome two-fold, rather than by silencing one of the female X chromosomes [190]. However, the coordination between lncRNAs and chromatin remodelers is again central in the process. RNA on the X (roX) lncRNAs are an integral component of the dosage compensation complex (DCC), and help to guide this complex to specific sites on the male X chromosome, which leads to the formation of euchromatin [191-194].

1.9 Summary of thesis aims and chapters

Without RNA interference, microRNAs, or many sequence-specific transcription factors, *P. falciparum* may rely heavily on transcriptional interference, epigenetic feedback, post-transcriptional modification, and nuclear positioning to regulate gene expression. In humans and model organisms, lncRNAs play dominant roles in all of the above mentioned gene regulatory mechanisms **[Table 1.1]** [143-145]. Moreover, the potential for lncRNAs to remain physically tethered at their site of origin, while interacting with other cellular factors such as chromatin remodeling complexes, provides these molecules with unique site-specific regulatory advantages [150]. Supporting this notion, lncRNAs have been found to play pivotal roles in allele-specific chromatin repression during XCI and imprinting [195], and site-specific chromatin repression during mating-type control of meiosis [181], and vernilization [186].

We hypothesized that lncRNAs may be important master regulators of *P. falciparum* gene expression and chromatin state as well. As of 2009, when I initiated this thesis work, there were preliminary reports describing the presence of developmentally regulated antisense RNA [196], *var* gene intron-derived RNA [140], and centromeric RNA in *P. falciparum* [197]. It was also clear that multi-gene families, such as the PfEMP1-encoding *var* genes, were epigenetically regulated in a monoallelic fashion, with intriguing parallels to lncRNA-mediated epigenetic regulation mechanisms emerging in humans and model organisms [106]. However, the functional significance of *P. falciparum* lncRNAs had not been investigated. A genome-wide map of *P. falciparum* lncRNAs was also fundamentally lacking, and lncRNA transcript properties, such as expression, conservation, and structure, had not been investigated on a broad-scale.

In **Chapter 2**, using a custom high-resolution DNA tiling array, we probe expression across approximately one quarter of the *P. falciparum* genome [198]. We identify blood stage-specific expression patterns for 60 putative *P. falciparum* lncRNAs and characterize a novel family of telomere-associated lncRNAs. Importantly, this represents the first experimental evidence showing that multiple *P. falciparum* chromosome ends coordinate transcription via a large non-coding RNA gene family, termed lncRNA-TARE (telomere-associated repetitive element). Our results indicate that lncRNA-TARE plays a central role in the transcriptional and epigenetic status of *P. falciparum* chromosome ends. Specifically, we find that all twenty-two lncRNA-TARE genes are located next to clinically important *var* genes; *var* genes are expressed through epigenetically mediated allelic exclusion [105]. We also find that lncRNA-TARE loci comprise the majority of subtelomeric *var* promoter element (SPE2) binding sites for the *P. falciparum* SPE2-interacting protein (PfSip2). PfSip2 is an ApiAP2 DNA-binding factor implicated in subtelomeric heterochromatin formation and *var* gene regulation [75].

In **Chapter 3**, we develop and validate strand-specific, non-polyA-selected RNA sequencing methods sensitive to the extreme AT content of *P. falciparum.* We use this approach to sequence the *P. falciparum* blood stage transcriptome across 56 hours of development at unprecedented resolution, including during red blood cell rupture and parasite invasion. We then pursue the first assembly of *P. falciparum* transcript structures from RNA sequencing data, and assemble over one thousand high-confidence, bona fide lncRNA transcripts. We perform extensive global analysis of 660 intergenic lncRNA and 474 antisense RNA transcript properties: coding prediction, periodicity, stage-specificity, correlation, GC content, length, splicing, and location relative to annotated transcripts. Our results indicate a functional role for many *P. falciparum* lncRNAs. For example, we find evidence of transcriptional interference between

sense-antisense partners, a subset of antisense RNAs whose transcript levels drop significantly during invasion, a global and a highly coordinated invasion-specific transcript profile for lncRNA-TAREs and 'sterile' *var* transcripts, and a multi-exonic and highly correlated antisense RNA counterpart to a gene required for sexual commitment. Lastly, we validate six *P. falciparum* circular RNA (circRNA) transcript structures in an organism lacking RNA interference, and predict the non-canonical splice junctions of over one thousand additional *P. falciparum* circRNAs.

In **Chapter 4**, I summarize the contributions of my thesis work, and suggest future directions for genome-wide and case-by-case *P. falciparum* lncRNA studies. Taken together, this thesis represents a node connecting recent advancements in the lncRNA and genomics fields to the biology of the devastating human pathogen *P. falciparum* malaria. At the unique intersection between these worlds, I develop novel high-throughput methods and discover compelling *P. falciparum* lncRNAs with predicted roles in development, virulence, and transmission.

1.10 References

1. World Health Organization (WHO) World Malaria Report 2013 http://www.who.int/malaria/publications/world_malaria_report_2013
2. Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, et al. (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet 379: 413-431.
3. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415: 673-679.
4. Miller LH, Ackerman HC, Su XZ, Wellems TE (2013) Malaria biology and disease pathogenesis: insights for new treatments. Nat Med 19: 156-167.
5. Tanner M, de Savigny D (2008) Malaria eradication back on the table. Bulletin of the World Health Organization 86: 82.
6. Steketee RW, Nahlen BL, Parise ME, Menendez C (2001) The burden of malaria in pregnancy in malaria-endemic areas. Am J Trop Med Hyg 64: 28-35.
7. Doolan DL, Dobano C, Baird JK (2009) Acquired immunity to malaria. Clin Microbiol Rev 22: 13-36, Table of Contents.
8. Plowe CV, Alonso P, Hoffman SL (2009) The potential role of vaccines in the elimination of falciparum malaria and the eventual eradication of malaria. J Infect Dis 200: 1646-1649.
9. mal ERACGoV, Abdulla S, Agre P, Alonso PL, Arevalo-Herrera M, et al. (2011) A research agenda for malaria eradication: vaccines. PLoS medicine 8: e1000398.
10. Raj DK, Nixon CP, Nixon CE, Dvorin JD, DiPetrillo CG, et al. (2014) Antibodies to PfSEA-1 block parasite egress from RBCs and protect against malaria infection. Science 344: 871-877.
11. Bordbar B, Tuikue Ndam N, Renard E, Jafari-Guemouri S, Tavul L, et al. (2014) Genetic diversity of VAR2CSA ID1-DBL2Xb in worldwide Plasmodium falciparum populations: Impact on vaccine design for placental malaria. Infect Genet Evol 25: 81-92.
12. Epstein JE, Tewari K, Lyke KE, Sim BK, Billingsley PF, et al. (2011) Live attenuated malaria vaccine designed to protect through hepatic CD8(+) T cell immunity. Science 334: 475-480.
13. Sinden RE, Carter R, Drakeley C, Leroy D (2012) The biology of sexual development of Plasmodium: the design and implementation of transmission-blocking strategies. Malar J 11: 70.
14. Shimp RL, Jr., Rowe C, Reiter K, Chen B, Nguyen V, et al. (2013) Development of a Pfs25-EPA malaria transmission blocking vaccine as a chemically conjugated nanoparticle. Vaccine 31: 2954-2962.
15. Alonso PL, Brown G, Arevalo-Herrera M, Binka F, Chitnis C, et al. (2011) A research agenda to underpin malaria eradication. PLoS Med 8: e1000406.
16. Su X, Hayton K, Wellems TE (2007) Genetic linkage and association analyses for trait mapping in Plasmodium falciparum. Nat Rev Genet 8: 497-506.
17. Wirth DF (2002) Biological revelations. Nature 419: 495-496.
18. Whitten MM, Shiao SH, Levashina EA (2006) Mosquito midguts and malaria: cell biology, compartmentalization and immunology. Parasite Immunol 28: 121-130.
19. Smith RC, Vega-Rodriguez J, Jacobs-Lorena M (2014) The Plasmodium bottleneck: malaria parasite losses in the mosquito vector. Mem Inst Oswaldo Cruz 0.
20. Gerald N, Mahajan B, Kumar S (2011) Mitosis in the human malaria parasite Plasmodium falciparum. Eukaryot Cell 10: 474-482.
21. Baker DA (2010) Malaria gametocytogenesis. Mol Biochem Parasitol 172: 57-65.
22. Rosenberg R, Wirtz RA, Schneider I, Burge R (1990) An estimation of the number of malaria sporozoites ejected by a feeding mosquito. Trans R Soc Trop Med Hyg 84: 209-212.
23. Yuda M, Ishino T (2004) Liver invasion by malarial parasites--how do malarial parasites break through the host barrier? Cell Microbiol 6: 1119-1125.
24. Wright GJ, Rayner JC (2014) Plasmodium falciparum erythrocyte invasion: combining function with immune evasion. PLoS Pathog 10: e1003943.
25. Goldberg DE, Cowman AF (2010) Moving in and renovating: exporting proteins from Plasmodium into host erythrocytes. Nature Reviews Microbiology 8: 617-621.
26. Bannister LH, Hopkins JM, Fowler RE, Krishna S, Mitchell GH (2000) A brief illustrated guide to the ultrastructure of Plasmodium falciparum asexual blood stages. Parasitol Today 16: 427-433.
27. Maier AG, Cooke BM, Cowman AF, Tilley L (2009) Malaria parasite proteins that remodel the host erythrocyte. Nature Reviews Microbiology 7: 341-354.

28. Bruce MC, Alano P, Duthie S, Carter R (1990) Commitment of the malaria parasite Plasmodium falciparum to sexual and asexual development. Parasitology 100 Pt 2: 191-200.
29. Taylor LH, Read AF (1997) Why so few transmission stages? Reproductive restraint by malaria parasites. Parasitology today (Personal ed) 13: 135-140.
30. Silvestrini F, Alano P, Williams JL (2000) Commitment to the production of male and female gametocytes in the human malaria parasite Plasmodium falciparum. Parasitology 121 Pt 5: 465-471.
31. Talman AM, Domarle O, McKenzie FE, Ariey F, Robert V (2004) Gametocytogenesis: the puberty of Plasmodium falciparum. Malar J 3: 24.
32. Hawking F, Wilson ME, Gammage K (1971) Evidence for cyclic development and short-lived maturity in the gametocytes of Plasmodium falciparum. Trans R Soc Trop Med Hyg 65: 549-559.
33. Bruce MC, Carter RN, Nakamura K, Aikawa M, Carter R (1994) Cellular location and temporal expression of the Plasmodium falciparum sexual stage antigen Pfs16. Molecular and biochemical parasitology 65: 11-22.
34. Carter R, Graves PM, Creasey A, Byrne K, Read D, et al. (1989) Plasmodium falciparum: an abundant stage-specific protein expressed during early gametocyte development. Experimental parasitology 69: 140-149.
35. Ingmundson A, Alano P, Matuschewski K, Silvestrini F (2014) Feeling at home from arrival to departure: protein export and host cell remodelling during Plasmodium liver stage and gametocyte maturation. Cellular microbiology 16: 324-333.
36. Silvestrini F, Lasonder E, Olivieri A, Camarda G, van Schaijk B, et al. (2010) Protein export marks the early phase of gametocytogenesis of the human malaria parasite Plasmodium falciparum. Mol Cell Proteomics 9: 1437-1448.
37. Bousema T, Drakeley C (2011) Epidemiology and infectivity of Plasmodium falciparum and Plasmodium vivax gametocytes in relation to malaria control and elimination. Clin Microbiol Rev 24: 377-410.
38. Smalley ME, Abdalla S, Brown J (1981) The distribution of Plasmodium falciparum in the peripheral blood and bone marrow of Gambian children. Trans R Soc Trop Med Hyg 75: 103-105.
39. Dixon MW, Dearnley MK, Hanssen E, Gilberger T, Tilley L (2012) Shape-shifting gametocytes: how and why does P. falciparum go banana-shaped? Trends Parasitol 28: 471-478.
40. Marsh K, Forster D, Waruiru C, Mwangi I, Winstanley M, et al. (1995) Indicators of life-threatening malaria in African children. N Engl J Med 332: 1399-1404.
41. Reeder JC, Rogerson SJ, al-Yaman F, Anders RF, Coppel RL, et al. (1994) Diversity of agglutinating phenotype, cytoadherence, and rosette-forming characteristics of Plasmodium falciparum isolates from Papua New Guinean children. Am J Trop Med Hyg 51: 45-55.
42. Chotivanich K, Udomsangpetch R, Simpson JA, Newton P, Pukrittayakamee S, et al. (2000) Parasite multiplication potential and the severity of Falciparum malaria. J Infect Dis 181: 1206-1209.
43. Wendler JP, Okombo J, Amato R, Miotto O, Kiara SM, et al. (2014) A Genome Wide Association Study of Plasmodium falciparum Susceptibility to 22 Antimalarial Drugs in Kenya. PLoS One 9: e96486.
44. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, et al. (2007) A genome-wide map of diversity in Plasmodium falciparum. Nature genetics 39: 113-119.
45. Williams TN (2006) Human red blood cell polymorphisms and malaria. Curr Opin Microbiol 9: 388-394.
46. Williams TN, Mwangi TW, Wambua S, Peto TEA, Weatherall DJ, et al. (2005) Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. Nature genetics 37: 1253-1257.
47. Bunn HF (2013) The triumph of good over evil: protection by the sickle gene against malaria. Blood 121: 20-25.
48. Crabb BS, Cowman AF (2002) Plasmodium falciparum virulence determinants unveiled. Genome Biol 3: REVIEWS1031.
49. Beeson JG, Brown GV (2002) Pathogenesis of Plasmodium falciparum malaria: the roles of parasite adhesion and antigenic variation. Cell Mol Life Sci 59: 258-271.
50. Miller LH, Good MF, Milon G (1994) Malaria pathogenesis. Science 264: 1878-1883.
51. Langhi DM, Jr., Bordin JO (2006) Duffy blood group and malaria. Hematology 11: 389-398.
52. Gaur D, Storry JR, Reid ME, Barnwell JW, Miller LH (2003) Plasmodium falciparum is able to invade erythrocytes through a trypsin-resistant pathway independent of glycophorin B. Infection and immunity 71: 6742-6746.
53. Coleman BI, Ribacke U, Manary M, Bei AK, Winzeler EA, et al. (2012) Nuclear repositioning precedes promoter accessibility and is linked to the switching frequency of a Plasmodium falciparum invasion gene. Cell Host Microbe 12: 739-750.
54. Baum J, Maier AG, Good RT, Simpson KM, Cowman AF (2005) Invasion by P. falciparum merozoites suggests a hierarchy of molecular interactions. PLoS pathogens 1: e37.

55. Nery S, Deans A-M, Mosobo M, Marsh K, Rowe JA, et al. (2006) Expression of Plasmodium falciparum genes involved in erythrocyte invasion varies among isolates cultured directly from patients. Molecular and biochemical parasitology 149: 208-215.
56. Buffet PA, Safeukui I, Deplaine G, Brousse V, Prendki V, et al. (2011) The pathogenesis of Plasmodium falciparum malaria in humans: insights from splenic physiology. Blood 117: 381-392.
57. Flick K, Chen Q (2004) var genes, PfEMP1 and the human host. Mol Biochem Parasitol 134: 3-9.
58. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.
59. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu JC, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. Plos Biology 1: 85-100.
60. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, et al. (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res 14: 2308-2318.
61. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9: 3273-3297.
62. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Molecular biology of the cell 13: 1977-2000.
63. Baum J, Papenfuss AT, Mair GR, Janse CJ, Vlachou D, et al. (2009) Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. Nucleic Acids Res 37: 3788-3798.
64. Coulson RM, Hall N, Ouzounis CA (2004) Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. Genome Res 14: 1548-1554.
65. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M (2010) Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. PLoS Pathog 6: e1001165.
66. Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res 33: 3994-4006.
67. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, et al. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. Proceedings of the National Academy of Sciences of the United States of America 105: 8393-8398.
68. Lai WS, Carballo E, Thorn JM, Kennington EA, Blackshear PJ (2000) Interactions of CCCH zinc finger proteins with mRNA. Binding of tristetraprolin-related zinc finger proteins to Au-rich elements and destabilization of mRNA. The Journal of biological chemistry 275: 17827-17837.
69. Iwanaga S, Kaneko I, Kato T, Yuda M (2012) Identification of an AP2-family protein that is critical for malaria liver stage development. PloS one 7: e47557.
70. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, et al. (2014) A cascade of DNA-binding proteins for sexual commitment and development in Plasmodium. Nature 507: 253-257.
71. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites. Molecular microbiology 75: 854-863.
72. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, et al. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Molecular microbiology 71: 1402-1414.
73. Kafsack BFC, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, et al. (2014) A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature 507: 248-252.
74. Zhang Q, Huang Y, Zhang Y, Fang X, Claes A, et al. (2011) A critical role of perinuclear filamentous actin in spatial repositioning and mutually exclusive expression of virulence genes in malaria parasites. Cell host & microbe 10: 451-463.
75. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BT, et al. (2010) A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. PLoS Pathog 6: e1000784.
76. Cui L, Miao J (2010) Chromatin-mediated Epigenetic Regulation in the Malaria Parasite Plasmodium falciparum. Eukaryot Cell.
77. Trelle MB, Salcedo-Amaya AM, Cohen AM, Stunnenberg HG, Jensen ON (2009) Global histone analysis by mass spectrometry reveals a high content of acetylated lysine residues in the malaria parasite Plasmodium falciparum. Journal of proteome research 8: 3439-3450.
78. Gupta AP, Chin WH, Zhu L, Mok S, Luah YH, et al. (2013) Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum. PLoS Pathog 9: e1003170.

79. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, et al. (2010) H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. Plos Pathogens 6: e1001223.
80. Salcedo-Amaya AM, van Driel MA, Alako BT, Trelle MB, van den Elzen AMG, et al. (2009) Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 106: 9655-9660.
81. Flueck C, Bartfai R, Volz J, Niederwieser I, Salcedo-Amaya AM, et al. (2009) Plasmodium falciparum Heterochromatin Protein 1 Marks Genomic Loci Linked to Phenotypic Variation of Exported Virulence Factors. Plos Pathogens 5: -.
82. Lopez-Rubio JJ, Mancio-Silva L, Scherf A (2009) Genome-wide Analysis of Heterochromatin Associates Clonally Variant Gene Regulation with Perinuclear Repressive Centers in Malaria Parasites. Cell Host & Microbe 5: 179-190.
83. Hoeijmakers WAM, Stunnenberg HG, Bartfai R (2012) Placing the Plasmodium falciparum epigenome on the map. Trends in parasitology 28: 486-495.
84. Painter HJ, Campbell TL, Llinas M (2011) The Apicomplexan AP2 family: integral factors regulating Plasmodium development. Molecular and Biochemical Parasitology 176: 1-7.
85. Hakimi M-A, Deitsch KW (2007) Epigenetics in Apicomplexa: control of gene expression during cell cycle progression, differentiation and antigenic variation. Current opinion in microbiology 10: 357-362.
86. Cortes A, Crowley VM, Vaquero A, Voss TS (2012) A view on the role of epigenetics in the biology of malaria parasites. PLoS Pathog 8: e1002943.
87. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, et al. (2013) Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum. Genome Biol 14: R128.
88. Shock JL, Fischer KF, DeRisi JL (2007) Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. Genome biology 8: R134.
89. Sims JS, Militello KT, Sims PA, Patel VP, Kasper JM, et al. (2009) Patterns of gene-specific and total transcriptional activity during the Plasmodium falciparum intraerythrocytic developmental cycle. Eukaryot Cell 8: 327-338.
90. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z (2008) Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites. Genome Biol 9: R177.
91. Dechering KJ, Thompson J, Dodemont HJ, Eling W, Konings RN (1997) Developmentally regulated expression of pfs16, a marker for sexual differentiation of the human malaria parasite Plasmodium falciparum. Molecular and biochemical parasitology 89: 235-244.
92. Vervenne RA, Dirks RW, Ramesar J, Waters AP, Janse CJ (1994) Differential expression in blood stages of the gene coding for the 21-kilodalton surface protein of ookinetes of Plasmodium berghei as detected by RNA in situ hybridisation. Molecular and biochemical parasitology 68: 259-266.
93. Rehkopf DH, Gillespie DE, Harrell MI, Feagin JE (2000) Transcriptional mapping and RNA processing of the Plasmodium falciparum mitochondrial mRNAs. Molecular and biochemical parasitology 105: 91-103.
94. Mair GR, Braks JAM, Garver LS, Wiegant JCAG, Hall N, et al. (2006) Regulation of sexual development of Plasmodium by translational repression. Science (New York, N Y) 313: 667-669.
95. Raabe AC, Billker O, Vial HJ, Wengelnik K (2009) Quantitative assessment of DNA replication to monitor microgametogenesis in Plasmodium berghei. Molecular and biochemical parasitology 168: 172-176.
96. Canning EU, Sinden RE (1975) Nuclear organisation in gametocytes of Plasmodium and Hepatocystis: a cytochemical study. Zeitschrift fur Parasitenkunde (Berlin, Germany) 46: 297-299.
97. Sinden RE, Canning EU, Bray RS, Smalley ME (1978) Gametocyte and gamete development in Plasmodium falciparum. Proceedings of the Royal Society of London Series B, Containing papers of a Biological character Royal Society (Great Britain) 201: 375-399.
98. Mair GR, Lasonder E, Garver LS, Franke-Fayard BMD, Carret CK, et al. (2010) Universal features of post-transcriptional gene regulation are critical for Plasmodium zygote development. PLoS pathogens 6: e1000767.
99. Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, et al. (2005) A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science (New York, N Y) 307: 82-86.
100. Miao J, Li J, Fan Q, Li X, Li X, et al. (2010) The Puf-family RNA-binding protein PfPuf2 regulates sexual development and sex differentiation in the malaria parasite Plasmodium falciparum. Journal of cell science 123: 1039-1049.

101. Wickens M, Bernstein DS, Kimble J, Parker R (2002) A PUF family portrait: 3'UTR regulation as a way of life. Trends in genetics : TIG 18: 150-157.
102. Pal-Bhowmick I, Vora HK, Jarori GK (2007) Sub-cellular localization and post-translational modifications of the Plasmodium yoelii enolase suggest moonlighting functions. Malaria journal 6: 45.
103. Zhang M, Mishra S, Sakthivel R, Rojas M, Ranjan R, et al. (2012) PK4, a eukaryotic initiation factor 2alpha(eIF2alpha) kinase, is essential for the development of the erythrocytic cycle of Plasmodium. Proceedings of the National Academy of Sciences of the United States of America 109: 3956-3961.
104. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, et al. (2012) Transcriptional variation in the malaria parasite Plasmodium falciparum. Genome Res 22: 925-938.
105. Guizetti J, Scherf A (2013) Silence, activate, poise and switch! Mechanisms of antigenic variation in Plasmodium falciparum. Cell Microbiol 15: 718-726.
106. Scherf A, Lopez-Rubio JJ, Riviere L (2008) Antigenic variation in Plasmodium falciparum. Annu Rev Microbiol 62: 445-470.
107. Gatton ML, Peters JM, Fowler EV, Cheng Q (2003) Switching rates of Plasmodium falciparum var genes: faster than we thought? Trends in parasitology 19: 202-208.
108. Comeaux CA, Coleman BI, Bei AK, Whitehurst N, Duraisingh MT (2011) Functional analysis of epigenetic regulation of tandem RhopH1/clag genes reveals a role in Plasmodium falciparum growth. Molecular Microbiology 80: 378-390.
109. Cortes A, Carret C, Kaneko O, Yim Lim BY, Ivens A, et al. (2007) Epigenetic silencing of Plasmodium falciparum genes linked to erythrocyte invasion. PLoS Pathog 3: e107.
110. Pillai AD, Nguitragool W, Lyko B, Dolinta K, Butler MM, et al. (2012) Solute restriction reveals an essential role for clag3-associated channels in malaria parasite nutrient acquisition. Mol Pharmacol 82: 1104-1114.
111. Nguitragool W, Bokhari AAB, Pillai AD, Rayavara K, Sharma P, et al. (2011) Malaria parasite clag3 genes determine channel-mediated nutrient uptake by infected red blood cells. Cell 145: 665-677.
112. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, et al. (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. Cell 121: 25-36.
113. Volz JC, Bartfai R, Petter M, Langer C, Josling GA, et al. (2012) PfSET10, a Plasmodium falciparum methyltransferase, maintains the active var gene in a poised state during parasite division. Cell host & microbe 11: 7-18.
114. Avraham I, Schreier J, Dzikowski R (2012) Insulator-like pairing elements regulate silencing and mutually exclusive expression in the malaria parasite Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America 109: E3678-3686.
115. Deitsch KW, Calderwood MS, Wellems TE (2001) Malaria. Cooperative silencing elements in var genes. Nature 412: 875-876.
116. Frank M, Dzikowski R, Costantini D, Amulic B, Berdougo E, et al. (2006) Strict pairing of var promoters and introns is required for var gene silencing in the malaria parasite Plasmodium falciparum. The Journal of biological chemistry 281: 9942-9952.
117. Swamy L, Amulic B, Deitsch KW (2011) Plasmodium falciparum var gene silencing is determined by cis DNA elements that form stable and heritable interactions. Eukaryotic cell 10: 530-539.
118. Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT, et al. (2005) Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium falciparum. Cell 121: 13-24.
119. Ralph SA, Scheidig-Benatar C, Scherf A (2005) Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations. Proc Natl Acad Sci U S A 102: 5414-5419.
120. Llinas M, Bozdech Z, Wong ED, Adai AT, DeRisi JL (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. Nucleic acids research 34: 1166-1173.
121. Zhou Y, Ramachandran V, Kumar KA, Westenberger S, Refour P, et al. (2008) Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. PloS one 3: e1570.
122. Ganesan K, Ponmee N, Jiang L, Fowble JW, White J, et al. (2008) A genetically hard-wired metabolic transcriptome in Plasmodium falciparum fails to mount protective responses to lethal antifolates. PLoS pathogens 4: e1000214.
123. Gunasekera AM, Myrick A, Le Roch K, Winzeler E, Wirth DF (2007) Plasmodium falciparum: genome wide perturbations in transcript profiles among mixed stage cultures after chloroquine treatment. Experimental parasitology 117: 87-92.

124. Chaal BK, Gupta AP, Wastuwidyaningtyas BD, Luah YH, Bozdech Z (2010) Histone Deacetylases Play a Major Role in the Transcriptional Regulation of the Plasmodium falciparum Life Cycle. Plos Pathogens 6: -.
125. Le Roch KG, Johnson JR, Ahiboh H, Chung D-WD, Prudhomme J, et al. (2008) A systematic approach to understand the mechanism of action of the bisthiazolium compound T4 on the human malaria parasite, Plasmodium falciparum. BMC genomics 9: 513.
126. Natalang O, Bischoff E, Deplaine G, Proux C, Dillies M-A, et al. (2008) Dynamic RNA profiling in Plasmodium falciparum synchronized blood stages exposed to lethal doses of artesunate. BMC genomics 9: 388.
127. Oakley MSM, Kumar S, Anantharaman V, Zheng H, Mahajan B, et al. (2007) Molecular factors and biochemical pathways induced by febrile temperature in intraerythrocytic Plasmodium falciparum parasites. Infection and immunity 75: 2012-2025.
128. Torrentino-Madamet M, Almeras L, Desplans J, Le Priol Y, Belghazi M, et al. (2011) Global response of Plasmodium falciparum to hyperoxia: a combined transcriptomic and proteomic approach. Malaria journal 10: 4.
129. Kato N, Sakata T, Breton G, Le Roch KG, Nagle A, et al. (2008) Gene expression signatures and small-molecule compounds link a protein kinase to Plasmodium falciparum motility. Nature chemical biology 4: 347-356.
130. Dyer M, Day KP (2000) Commitment to gametocytogenesis in Plasmodium falciparum. Parasitology today (Personal ed) 16: 102-107.
131. LaMonte G, Philip N, Reardon J, Lacsina JR, Majoros W, et al. (2012) Translocation of sickle cell erythrocyte microRNAs into Plasmodium falciparum inhibits parasite translation and contributes to malaria resistance. Cell host & microbe 12: 187-199.
132. Mantel P-Y, Hoang AN, Goldowitz I, Potashnikova D, Hamza B, et al. (2013) Malaria-infected erythrocyte-derived microvesicles mediate cellular communication within the parasite population and with the host immune system. Cell host & microbe 13: 521-534.
133. Regev-Rudzki N, Wilson DW, Carvalho TG, Sisquella X, Coleman BM, et al. (2013) Cell-cell communication between malaria-infected red blood cells via exosome-like vesicles. Cell 153: 1120-1133.
134. Eksi S, Morahan BJ, Haile Y, Furuya T, Jiang H, et al. (2012) Plasmodium falciparum gametocyte development 1 (Pfgdv1) and gametocytogenesis early gene identification and commitment to sexual development. PLoS Pathog 8: e1002964.
135. Bourke PF, Holt DC, Sutherland CJ, Kemp DJ (1996) Disruption of a novel open reading frame of Plasmodium falciparum chromosome 9 by subtelomeric and internal deletions can lead to loss or maintenance of cytoadherence. Molecular and Biochemical Parasitology 82: 25-36.
136. Spielmann T, Hawthorne PL, Dixon MWA, Hannemann M, Klotz K, et al. (2006) A cluster of ring stage-specific genes linked to a locus implicated in cytoadherence in Plasmodium falciparum codes for PEXEL-negative and PEXEL-positive proteins exported into the host cell. Molecular Biology of the Cell 17: 3613-3624.
137. Mackinnon MJ, Li J, Mok S, Kortok MM, Marsh K, et al. (2009) Comparative transcriptional and genomic analysis of Plasmodium falciparum field isolates. PLoS pathogens 5: e1000644.
138. Ikadai H, Shaw Saliba K, Kanzok SM, McLean KJ, Tanaka TQ, et al. (2013) Transposon mutagenesis identifies genes essential for Plasmodium falciparum gametocytogenesis. Proceedings of the National Academy of Sciences of the United States of America 110: E1676-1684.
139. Kim K (2012) Malaria var gene expression: keeping up with the neighbors. Cell host & microbe 11: 1-2.
140. Kyes SA, Christodoulou Z, Raza A, Horrocks P, Pinches R, et al. (2003) A well-conserved Plasmodium falciparum var gene shows an unusual stage-specific transcript pattern. Molecular microbiology 48: 1339-1348.
141. Epp C, Li F, Howitt CA, Chookajorn T, Deitsch KW (2009) Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. RNA 15: 116-127.
142. Lee JT (2012) Epigenetic regulation by long noncoding RNAs. Science (New York, N Y) 338: 1435-1439.
143. Geisler S, Coller J (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. Nat Rev Mol Cell Biol 14: 699-712.
144. Kung JTY, Colognori D, Lee JT (2013) Long noncoding RNAs: past, present, and future. Genetics 193: 651-669.

145. Modarresi F, Faghihi MA, Lopez-Toledano MA, Fatemi RP, Magistri M, et al. (2012) Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. Nature biotechnology 30: 453-459.
146. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, et al. (1992) The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71: 515-526.
147. Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, et al. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature 349: 38-44.
148. Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, et al. (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71: 527-542.
149. Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. Nature 379: 131-137.
150. Lee JT (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev 23: 1831-1842.
151. Marahrens Y, Panning B, Dausman J, Strauss W, Jaenisch R (1997) Xist-deficient mice are defective in dosage compensation but not spermatogenesis. Genes & development 11: 156-166.
152. Froberg JE, Yang L, Lee JT (2013) Guided by RNAs: X-inactivation as a model for lncRNA function. Journal of molecular biology 425: 3698-3706.
153. Lee JT (2011) Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. Nature reviews Molecular cell biology 12: 815-826.
154. Zhang L-F, Huynh KD, Lee JT (2007) Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. Cell 129: 693-706.
155. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, et al. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell 40: 939-953.
156. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. Science 322: 750-756.
157. Lee JT, Davidow LS, Warshawsky D (1999) Tsix, a gene antisense to Xist at the X-inactivation centre. Nature genetics 21: 400-404.
158. Sado T, Hoki Y, Sasaki H (2005) Tsix silences Xist through modification of chromatin structure. Developmental cell 9: 159-165.
159. Tian D, Sun S, Lee JT (2010) The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell 143: 390-403.
160. Ogawa Y, Lee JT (2003) Xite, X-inactivation intergenic transcription elements that regulate the probability of choice. Molecular cell 11: 731-743.
161. Donohoe ME, Silva SS, Pinter SF, Xu N, Lee JT (2009) The pluripotency factor Oct4 interacts with Ctcf and also controls X-chromosome pairing and counting. Nature 460: 128-132.
162. Navarro P, Chambers I, Karwacki-Neisius V, Chureau C, Morey C, et al. (2008) Molecular coupling of Xist regulation and pluripotency. Science (New York, N Y) 321: 1693-1695.
163. Hawkins PG, Morris KV (2010) Transcriptional regulation of Oct4 by a long non-coding RNA antisense to Oct4-pseudogene 5. Transcription 1: 165-175.
164. Edwards CA, Ferguson-Smith AC (2007) Mechanisms regulating imprinted genes in clusters. Current opinion in cell biology 19: 281-289.
165. Wan L-B, Bartolomei MS (2008) Regulation of imprinting in clusters: noncoding RNAs versus insulators. Advances in genetics 61: 207-223.
166. Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, et al. (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature 389: 745-749.
167. Latos PA, Pauler FM, Koerner MV, Senergin HB, Hudson QJ, et al. (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. Science 338: 1469-1472.
168. Lee MP, DeBaun MR, Mitsuya K, Galonek HL, Brandenburg S, et al. (1999) Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. Proceedings of the National Academy of Sciences of the United States of America 96: 5203-5208.
169. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, et al. (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol Cell 32: 232-246.

170. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, et al. (2008) Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nature medicine 14: 723-730.
171. Scheele C, Petrovic N, Faghihi MA, Lassmann T, Fredriksson K, et al. (2007) The human PINK1 locus is regulated in vivo by a non-coding natural antisense RNA during modulation of mitochondrial function. BMC genomics 8: 74.
172. Khalil AM, Faghihi MA, Modarresi F, Brothers SP, Wahlestedt C (2008) A novel RNA transcript with antiapoptotic function is silenced in fragile X syndrome. PloS one 3: e1486.
173. Cheetham SW, Gruhl F, Mattick JS, Dinger ME (2013) Long noncoding RNAs and the genetics of cancer. British journal of cancer 108: 2419-2425.
174. Wahlestedt C (2013) Targeting long non-coding RNA to therapeutically upregulate gene expression. Nature reviews Drug discovery 12: 433-446.
175. Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, et al. (2009) RNAi in budding yeast. Science (New York, N Y) 326: 544-550.
176. Nicolas FE, Torres-Martinez S, Ruiz-Vazquez RM (2013) Loss and retention of RNA interference in fungi and parasites. PLoS pathogens 9: e1003089.
177. Harrison BR, Yazgan O, Krebs JE (2009) Life without RNAi: noncoding RNAs and their functions in Saccharomyces cerevisiae. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire 87: 767-779.
178. Hainer SJ, Pruneski JA, Mitchell RD, Monteverde RM, Martens JA (2011) Intergenic transcription causes repression by directing nucleosome assembly. Genes Dev 25: 29-40.
179. Hongay CF, Grisafi PL, Galitski T, Fink GR (2006) Antisense transcription controls cell fate in Saccharomyces cerevisiae. Cell 127: 735-745.
180. van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, et al. (2012) Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. Cell 150: 1170-1181.
181. van Werven FJ, Amon A (2011) Regulation of entry into gametogenesis. Philosophical transactions of the Royal Society of London Series B, Biological sciences 366: 3521-3531.
182. Houseley J, Rubbi L, Grunstein M, Tollervey D, Vogelauer M (2008) A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. Mol Cell 32: 685-695.
183. Bumgarner SL, Dowell RD, Grisafi P, Gifford DK, Fink GR (2009) Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast. Proceedings of the National Academy of Sciences of the United States of America 106: 18321-18326.
184. Bumgarner SL, Neuert G, Voight BF, Symbor-Nagrabska A, Grisafi P, et al. (2012) Single-cell analysis reveals that noncoding RNAs contribute to clonal heterogeneity by modulating transcription factor recruitment. Molecular cell 45: 470-482.
185. Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F (2007) Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae. Cell 131: 706-717.
186. Heo JB, Lee Y-S, Sung S (2013) Epigenetic regulation by long noncoding RNAs in plants. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 21: 685-693.
187. Au PCK, Zhu Q-H, Dennis ES, Wang M-B (2011) Long non-coding RNA-mediated mechanisms independent of the RNAi pathway in animals and plants. RNA biology 8: 404-414.
188. Song J, Irwin J, Dean C (2013) Remembering the prolonged cold of winter. Current biology : CB 23: R807-811.
189. Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76-79.
190. Gelbart ME, Kuroda MI (2009) Drosophila dosage compensation: a complex voyage to the X chromosome. Development (Cambridge, England) 136: 1399-1410.
191. Amrein H, Axel R (1997) Genes expressed in neurons of adult male Drosophila. Cell 88: 459-469.
192. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Molecular cell 44: 667-678.
193. Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL (1997) roX1 RNA paints the X chromosome of male Drosophila and is regulated by the dosage compensation system. Cell 88: 445-457.
194. Simon MD, Wang CI, Kharchenko PV, West JA, Chapman BA, et al. (2011) The genomic binding sites of a noncoding RNA. Proceedings of the National Academy of Sciences of the United States of America 108: 20497-20502.
195. Lee JT (2012) Epigenetic regulation by long noncoding RNAs. Science 338: 1435-1439.

196. Militello KT, Patel V, Chessler A-D, Fisher JK, Kasper JM, et al. (2005) RNA polymerase II synthesizes antisense RNA in Plasmodium falciparum. RNA (New York, N Y) 11: 365-370.
197. Li F, Sonbuchner L, Kyes SA, Epp C, Deitsch KW (2008) Nuclear non-coding RNAs are transcribed from the centromeres of Plasmodium falciparum and are associated with centromeric chromatin. The Journal of biological chemistry 283: 5692-5698.
198. Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, et al. (2011) A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. Genome biology 12: R56.

Chapter 2 | **A global transcriptional analysis of *Plasmodium falciparum* malaria reveals a novel family of telomere-associated lncRNAs**

Contributions: Kate M Broadbent[1,2], Daniel Park[2,3,], Ashley R. Wolf[1,2], Daria Van Tyne[4], Jennifer S. Sims[4], Ulf Ribacke[4], Sarah Volkman[2,4,5], Manoj Duraisingh[4], Dyann Wirth[2,4], Pardis C. Sabeti*[1,2,3,6§], John L. Rinn*[1,2,7,8§] . (*) equal contribution.

This chapter originally appeared as Broadbent et al. *Genome Biology*, 2011 [1].

Abstract

Mounting evidence suggests a major role for epigenetic feedback in *Plasmodium falciparum* transcriptional regulation. Long non-coding RNAs (lncRNAs) have recently emerged as a new paradigm in epigenetic remodeling. We therefore set out to investigate putative roles for lncRNAs in *P. falciparum* transcriptional regulation. We used a high-resolution DNA tiling microarray to survey transcriptional activity across 22.6% of the *P. falciparum* strain 3D7 genome. We identified 872 protein-coding genes and 60 putative *P. falciparum* lncRNAs under developmental regulation during the parasite's pathogenic human blood stage. Further characterization of lncRNA candidates led to the discovery of an intriguing family of long non-coding RNA telomere-associated repetitive element transcripts, termed lncRNA-TARE. We have quantified lncRNA-TARE expression at fifteen distinct chromosome ends and mapped putative transcriptional start and termination sites of lncRNA-TARE loci. Remarkably, we observed coordinated and stage-specific expression of lncRNA-TARE on all chromosome ends tested, and

two dominant transcripts of approximately 1.5 kb and 3.1 kb transcribed towards the telomere. In conclusion, we have characterized a family of twenty-two telomere-associated lncRNAs in *P. falciparum*. Homologous lncRNA-TARE loci are coordinately expressed after parasite DNA replication, and are poised to play an important role in *P. falciparum* telomere maintenance, virulence gene regulation, and potentially other processes of parasite chromosome end biology. Further study of lncRNA-TARE and other promising lncRNA candidates may provide mechanistic insight into *P. falciparum* transcriptional regulation.

Authors' contributions

JLR and PCS conceived of the study and participated in its design, coordination, and interpretation. DP carried out the computational design of arrays. JS harvested and extracted samples for arrays, and KMB harvested and extracted samples for qRT-PCR and RLM-RACE. ARW performed sample labeling and array hybridizations. KMB analyzed raw and normalized data and wrote/implemented the sliding-window TAR detection algorithm. KMB and DP analyzed predicted TARs to assess biological validity and filter for lncRNA candidates. KMB and DP characterized putative lncRNAs, and KMB designed/performed qRT-PCR and RLM-RACE analysis. DW, MD, SV, DVT, and UR contributed to the acquisition and interpretation of data. KMB wrote the manuscript, JLR and PCS critically revised the manuscript, and all authors have read and given approval of the version to be published.

Acknowledgements

We would like to thank M. Koziol and D. Shechner of the Broad Institute and Harvard University for experimental assistance and for Figure 3b illustration support, respectively. We would also like to thank members of the Broad Institute-Harvard Malaria Initiative for helpful feedback throughout. This work was supported by grants from the Gates Foundation and the National Institutes of Health Director's New Innovator Awards (1DP2OD00667-01). P.C.S. is a Packard and Burroughs Wellcome fellow. J.L.R is a Damon Runyon-Rachleff, Searle, Smith Family and Merkin Fellow. K.B and D.P are National Science Foundation Graduate Research Fellows.

2.1 Introduction

The causative agent of the most severe form of human malaria, *P. falciparum*, is a unicellular eukaryotic parasite transmitted through the bites of infected mosquitoes. The most vulnerable population to malarial disease is African children, but a staggering 3.3 billion people - half the world's population – are at risk for malarial infection. Despite recent research advances [2-7], the mechanisms *P. falciparum* utilizes to regulate mutually exclusive expression of multi-gene virulence families and stage-specific expression of approximately 80% of its genome during pathogenic blood stage development remain elusive.

Most confounding is the scarcity of sequence-specific transcription factors and cis-acting regulatory elements, coupled with the apparent lack of both RNA interference machinery and DNA methylation in the parasite [8,9]. However, the recent discovery of an expanded lineage of 27 ApiAP2 transcription factors may partially explain how the parasite regulates its unusual genome [10,11]. Additionally, it is becoming increasingly clear that chromatin remodeling and epigenetic memory play an important role in blood stage-specific expression and antigenic variation of virulence genes [7,12,13]. Notably, while the parasite lacks many of the conventional regulatory mechanisms of other organisms, it has a full arsenal of conserved histone modifying enzymes, and a higher than average number of RNA-binding proteins [9,14].

In eukaryotes spanning from yeast to humans, epigenetic regulation incorporates feedback from non-coding RNAs. Specifically, long non-coding RNAs (lncRNAs) and small non-coding RNAs often interface with RNA binding proteins and chromatin remodeling complexes to modulate their targeted genomic loci [15-19]. For example, in X chromosome inactivation at least seven distinct lncRNAs coordinate the selection and silencing of an entire

chromosome [20]. As another example, long telomeric repeat-containing RNA (TERRA) transcripts have been recently discovered as a major constituent of telomeric heterochromatin. TERRA interacts with telomere-associated proteins such as telomerase, is developmentally regulated, and is implicated in telomere replication and structural maintenance processes [21-26].

To investigate putative regulatory roles for lncRNAs in *P. falciparum*, we designed a high-resolution DNA tiling array to survey transcriptional activity during the parasite's pathogenic human blood stage. We identified 60 lncRNA candidates and characterized their GC content, evolutionary conservation, expression profile, and correlation with neighboring genes. Notably, our transcriptional profiling and subsequent analysis revealed an outlier on all fronts: a long telomere-associated non-coding RNA gene, termed lncRNA-TARE-4L, encoded in the telomere-associated repetitive element (TARE) tract of chromosome four.

Upon further investigation of the lncRNA-TARE-4L locus, we discovered a multi-gene family of lncRNA-TAREs. We have mapped homologous lncRNA-TARE loci on twenty-two of twenty-eight *P. falciparum* chromosome ends, and quantified the coordinated, stage-specific transcription of fifteen distinct lncRNA-TARE sequences using quantitative Real-Time PCR (qRT-PCR). We additionally employed RNA Ligase Mediated Rapid Amplification of cDNA Ends (RLM-RACE) to map putative transcriptional start and termination sites of lncRNA-TARE genes, including three sequences not investigated by qRT-PCR. Our RLM-RACE results suggested two dominant transcripts of approximately 1.5 kb and 3.1 kb are transcribed from the TARE 3 boundary towards the telomere. Interestingly, we also found that an upsB-type *var* virulence gene is adjacent to each predicted lncRNA-TARE gene and that lncRNA-TARE sequence is enriched with transcription factor binding sites only otherwise found in upsB-type *var* gene promoters.

Our results complement the recent *P. falciparum* transcriptome studies of Otto et al., Raabe et al., and others by providing stage-specific profiling and characterization of several previously unidentified *P. falciparum* lncRNA candidates, including a long telomere-associated non-coding RNA family [6,27]. Specifically, we have demonstrated that long telomere-associated lncRNAs are coordinately expressed after parasite DNA replication from at least eighteen chromosome ends. Taken together, this work provides new insights into *P. falciparum* non-coding RNA transcription and contributes a previously uncharacterized parallel between *P. falciparum* and model eukaryote chromosome end biology.

2.2 Results

Tiling Microarray Experimental Design

In order to comprehensively identify and characterize long non-coding transcripts in *P. falciparum*, we selected overlapping probes tiling approximately 22.6% of the *P. falciparum* genome at 12 bp median resolution. Notably, our DNA tiling array design provides over one order of magnitude denser probe coverage than previous *P. falciparum* transcriptional profiling arrays, and is unique in that it deeply samples both genic and intergenic sequence [5]. Probes cover 561 Watson (+) strand protein-coding genes, 699 Crick (-) strand protein-coding genes, two ribosomal RNA genes, and 1.73 Mb of intergenic sequence on *P. falciparum* chromosomes 2, 3*, 4, 5*, 7, 9, 12* (* indicates partial coverage) [see **2.4 Methods** and **Supplementary Figure 1.1** for further genome coverage details].

We harvested RNA from highly synchronous 3D7 parasites to capture global transcriptional changes during the parasite's two major intraerythrocytic developmental cycle (IDC) transitions: ring to trophozoite and trophozoite to schizont stage. During the *P. falciparum*

IDC, the parasite first exports cytoadherence surface proteins to sequester itself in host tissue (ring stage = 0-24 hours). This is followed by hemoglobin digestion and DNA replication (trophozoite stage = 24-36 hours), and, finally, nuclei segmentation and formation of 16-32 daughter merozoites (schizont stage = 36-48) [28,29]. Specifically, we profiled the polyadenylated RNA population transcribed from both genomic strands at 18 (ring), 24 (ring/trophozoite), 30 (trophozoite), and 36 (trophozoite/schizont) +/- 3 hours post-erythrocyte invasion (hpi) using our custom DNA tiling microarray.

Identification of lncRNA candidates

To identify transcriptionally active regions (TARs) along the *P. falciparum* genome, we analyzed normalized probe hybridization intensities using a window-based scan statistic algorithm. As proposed by Guttman et al, a TAR can be reasonably defined as a contiguous stretch of tiling probes with signal intensity significantly above the background distribution of permuted hybridization intensities [30,31]. Briefly, we calculated a scan statistic score for iterative window intervals, controlling the probability of one or more intervals being erroneously called significant at 5% [32]. As the final step, we merged overlapping intervals to define TAR boundaries. Taken together, this highly conservative approach corrects for multiple testing and provides strong family-wise type I error rate (FWER) control in our set of significant TARs.

Our transcriptional profiling approach identified 1360 significantly expressed TARs. Specifically, 1229 TARs exhibited overlap with 872 probed *PlasmoDB v6.5* protein-coding genes, 8 TARs overlapped the two probed ribosomal RNA genes, and 123 TARs represented un-annotated, putative non-coding transcripts from *P. falciparum* intergenic regions [**Figure 2.1a**]. In summary, both ribosomal RNA genes and 64.1% of protein-coding genes on the array were

identified as expressed in at least one of the four time-points tested. This is consistent with previous studies showing that 70-90% of protein-coding genes are expressed during *P. falciparum* intraerythrocytic development [4-6].

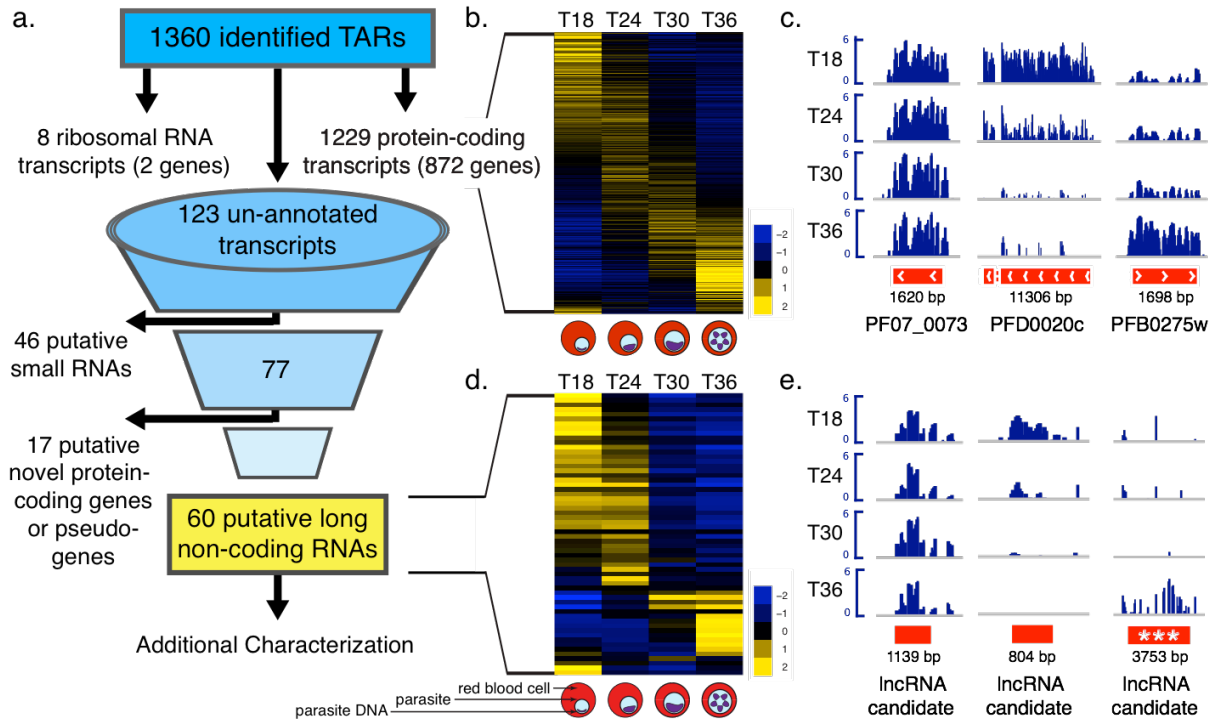


Figure 2.1 | Global transcriptional profiling of the *P. falciparum* IDC reveals 60 putative lncRNAs under developmental regulation. (a) We identified 1360 transcriptionally active regions (TARs) along the *P. falciparum* genome, of which 1229, 8, and 123 predicted transcripts exhibit overlap with known protein-coding genes, ribosomal RNA genes, and un-annotated intergenic regions, respectively. Further filtering of un-annotated transcripts for putative long non-coding RNAs (lncRNAs) eliminated 46 putative small RNAs and 17 putative novel protein-coding genes or pseudogenes, leaving a set of 60 putative lncRNAs for additional characterization. **(b/d)** Protein-coding and putative lncRNA transcripts follow the global expression cascade expected during the *P. falciparum* IDC. The mean log2 probe hybridization for each transcript is plotted from time-points T18, T24, T30, and T36 +/- 3 hpi corresponding to the ring, ring/trophozoite, trophozoite, and trophozoite/schizont intraerythrocytic stages of parasite growth (pictured below each column). Rows are mean centered. **(c/e)** Housekeeping gene seryl-tRNA synthetase (PF07_0073), early stage marker (PFD0020c), and late stage marker (PFB0275w) exhibit expected differential expression patterns. Examples of putative lncRNAs are also shown, including lncRNA-TARE-4L (***). Positive probe hybridization intensities are plotted from each sample after quantile normalization, log2 transformation, and median centering.

We next investigated our expression data for known patterns of gene regulation. We performed clustering analysis to confirm the stage-specific transcription cascade expected for *P. falciparum*'s protein-coding genome [**Figure 2.1b** and **Supplementary Figure 2.2**], and ontology analysis of stage-specific genes to confirm the parasite processes known to occur during our time-points [**Supplementary Table 2.1**]. As a further control, we confirmed that known housekeeping genes such as seryl-tRNA synthetase (PF07_0073) were not differentially expressed, while early and late stage markers such as PFD0020c and PFB0275w were maximally expressed in T18 and T36, respectively [**Figure 2.1c**] [4-6]. Collectively, these analyses confirmed the biological representation of our samples and provided meaningful context to investigate lncRNA expression during the parasite's pathogenic human blood stage.

We applied conservative criteria to identify bonafide *P. falciparum* lncRNAs from the set of 123 expressed intergenic transcripts. Namely, we required transcripts to be larger than 200 nucleotides, eliminating 46 putative small RNAs from further analysis [**Supplementary Table 2.2**]. We also eliminated transcripts having even marginal protein-coding potential. Briefly, we used BLASTX to translate the remaining 77 transcripts and search for significant protein matches across all 439,884 and 12,597,337 sequences represented in the Swissprot and non-redundant protein sequence (nr) databases, respectively. We also searched subsets of both databases with the following organism queries*: Plasmodium, Plasmodium falciparum,* and *Plasmodium falciparum strain 3D7*. While the large majority of analyzed transcripts lacked any coding potential, BLASTX analysis predicted that seventeen transcripts might, in fact, be novel *P. falciparum* genes or pseudogenes [**Supplementary Table 2.3**]. Thus our conservatively filtered set of putative lncRNAs for additional characterization consisted of 60 candidate sequences [**Figure 2.1a** and **Supplementary Table 2.4**].

Additional Characterization of lncRNA Candidates Highlights a Novel Telomere-Associated lncRNA

To systematically prioritize the 60 putative lncRNAs for functional follow-up, we looked for lncRNAs with similar properties to known functional transcripts in *P. falciparum* and/or bonafide lncRNAs in other organisms. We also investigated the possibility that lncRNA candidates may be spliced to nearby genes or represent un-annotated untranslated regions (UTRs). To this end, we assessed each lncRNA candidate's GC content and evolutionary conservation, as well as the correlation between lncRNA and neighboring gene expression profiles, and the distance between lncRNA and neighboring genes to infer transcript independence.

Known classes of functional non-coding RNA in *P. falciparum*, such as ribosomal RNA and transfer RNA, have high GC content [33,34]. While high GC content is certainly not sufficient or necessary for function, we were nonetheless interested in the GC content of our putative lncRNAs. We found the average GC content of lncRNA candidates (15.4%) to be typical of *P. falciparum* intergenic regions [35] and well below the coding transcript average (23.7%). This is not an unexpected result and, importantly, indicates no systematic hybridization bias towards detection of expressed non-coding regions with higher than expected GC content. Interestingly, however, this analysis highlighted one lncRNA with similar GC content (32.1%) to ribosomal RNA transcripts [**Figure 2.2a**]. We termed this candidate lncRNA-TARE-4L, as it is encoded in the telomere-associated repetitive element (TARE) tract on the left end of chromosome four.

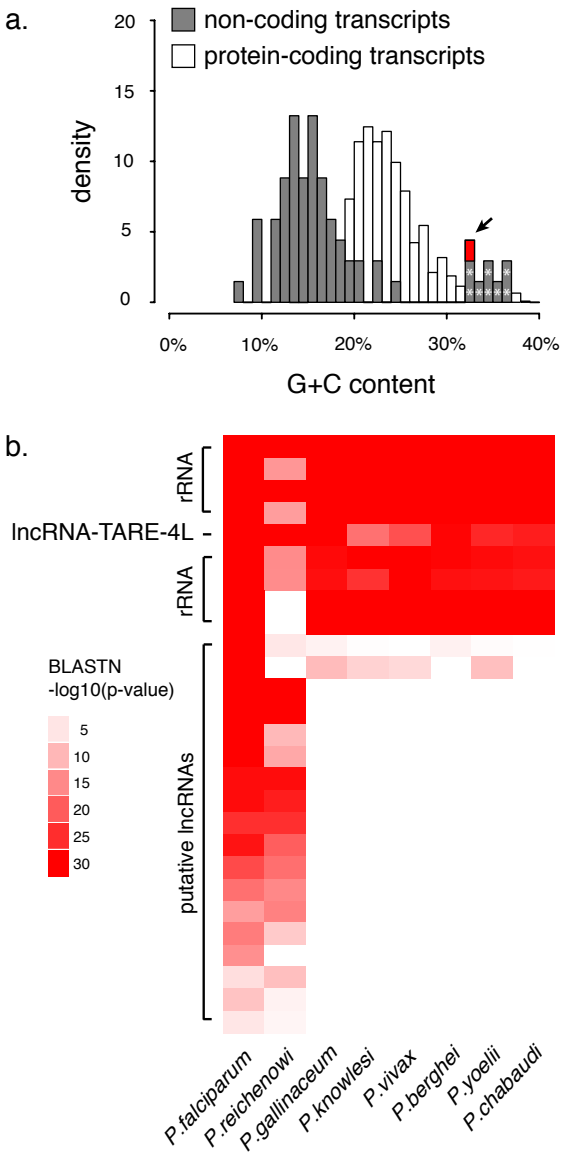


Figure 2.2 | Additional characterization of putative lncRNAs highlights the striking properties of a telomere-associated long non-coding RNA (a) GC content comparison between predicted non-coding and protein-coding transcripts shows that non-coding GC content is typically lower than coding transcript GC content. LncRNA-TARE-4L (red/arrow) and ribosomal RNAs (stars) have high GC content. **(b)** The sequence of eight functional ribosomal RNAs is broadly conserved across *Plasmodium* species. The sequence of nineteen putative lncRNAs is also significantly conserved in at least one other species (BLASTN p-value < 0.01), suggesting a preserved functional role. LncRNA-TARE-4L is strongly conserved across all eight species analyzed. Such a broad level of conservation is rarely seen in randomly sampled intergenic regions from *P. falciparum* (p-value = 0.045). Rows are ordered by number of species with BLASTN hits (p-value < 0.01), and then by strength of hits (defined by the product of p-values). Species are clustered based on the phylogeny proposed by Escalante and Alaya [86].

We next explored the evolutionary conservation of putative lncRNA sequences as an indicator of preserved functionality [36]. We performed BLASTN sequence alignment within the partially assembled, closely related *P. reichenowi* chimpanzee parasite genome and across six other partially sequenced, more distant *Plasmodium* species [**Figure 2.2b**]. We found that nineteen lncRNAs exhibited some level of conservation in the *Plasmodium* species investigated, but that only lncRNA-TARE-4L was strongly conserved across all species tested. To assess the significance of this result, we repeated this analysis for 600 size-matched random intergenic sequences, as well as for ribosomal RNAs and size-matched coding exons. We found that, respectively, 4.5%, 87.5%, and 33.3% had BLASTN hits to all *Plasmodium* species analyzed. Thus, the broad sequence conservation of lncRNA-TARE-4L is significant (p-value = 0.045) and more similar to the level of conservation expected for ribosomal RNAs and coding regions. Interestingly, we also found that 23 of the 27 broadly conserved null intergenic sequences map to either the telomeric or subtelomeric repeats.

The vast majority of *P. falciparum* genes are highly expressed only once per 48-hour IDC, and genes in related cellular processes are induced together [4]. Given this model for functional protein-coding transcripts, we reasoned that lncRNA candidates are more likely biologically significant if expressed in a stage-specific manner. Moreover, regulatory lncRNAs have been reported to act in cis to their targeted loci [36-38]. Consistent with this notion, lncRNAs that are both differentially expressed and encoded nearby essential or pathogenic genes may be involved in regulating these important loci. Figures 1d and 1e show that putative lncRNAs are developmentally regulated similar to protein-coding transcripts [See **2.4 Methods** and **Supplementary Figure 2.2** for detailed comparisons]. Additionally, we found that many lncRNA candidates neighbor essential genes and factors involved in parasite pathogenesis.

Given clear patterns of transcriptional regulation, we next tested if the 60 candidate lncRNAs were likely to be independent transcripts as opposed to un-annotated UTRs or small spliced exons to neighboring coding genes. To this end, we measured the Pearson correlation between putative lncRNA and neighboring gene expression profiles, conservatively selecting the most correlated, adjacent, expressed gene as the neighboring gene [**Supplementary Table 2.4** and **Methods 2.5**]. We found that 40 lncRNA candidates were highly correlated, raising the possibility that they may be spliced to or represent UTRs of adjacent coding genes. Of the remaining 20 lncRNA candidates exhibiting patterns of correlation consistent with independent transcription, 16 displayed biologically meaningful variation across our time-points. Thus, correlation analysis highlighted 16 candidates with clear evidence of independent transcriptional regulation.

While the null distribution of correlations from adjacent pairs of coding genes demonstrated that approximately 30 of 40 highly correlated lncRNA candidates may be UTRs or otherwise spliced to nearby genes, it also suggested that ten independent transcripts should exhibit this level of correlation by chance [**Supplemental Figure 2.3**]. To further distinguish such potential transcripts, we considered the distance of each putative lncRNA to the nearest neighboring coding gene. We found that 11 lncRNA candidates were over 1 kb from a coding gene. As the mean intron length in *P. falciparum* is 168 bp [35], we considered this as evidence of independence regardless of correlation value. Taken together, these results point to a highly conservative set of 23 lncRNA candidates based on correlation and distance analyses.

Strikingly, we found that lncRNA-TARE-4L exhibited the strongest evidence of independent transcriptional regulation. LncRNA-TARE-4L is flanked by the telomere and an upsB-type *var* gene [**Figure 2.3a**]. However, the *var* gene is over 20kb away and the exon

structure of *var* transcripts has been previously mapped [39,40]. Given a mean intron length of 168 bp and maximum intron length of 4.9 kb in *P. falciparum* [35], the splicing of lncRNA-TARE-4L to the nearest neighboring *var* gene would be biologically unprecedented. Moreover, correlation analysis revealed that lncRNA-TARE-4L is anti-correlated with the nearest expressed gene (r = -0.296). The nearest expressed gene (PFD0020c) is an additional 20 kb beyond the silenced upsB-type *var* gene and is profiled in Figure 1c for visual comparison to lncRNA-TARE-4L [Figure 1e, stars].

Collectively, lncRNA-TARE-4L emerged as the lncRNA candidate with the most promising properties for functional follow-up. LncRNA-TARE-4L is encoded in the TARE 2-3 subtelomeric repeat region between the telomere and a silenced upsB-type *var* gene [**Figure 2.3a**], has GC content and sequence conservation similar to that of functional ribosomal RNA, is sharply induced after parasite DNA replication, and is clearly an independent transcript.

As we further investigated the sequence properties of lncRNA-TARE-4L, we found that homologous lncRNA-TARE sequences are encoded adjacent to upsB-type *var* genes on twenty-two of twenty-eight *P. falciparum* chromosome ends [**Figure 2.3b**] [41-43]. In five cases where there is no lncRNA-TARE gene, there is similarly not an upsB-type *var* gene. This perhaps suggests concurrent evolutionary pressure acting on lncRNA-TARE and upsB-type *var* genes. Consistent with this notion, the LncRNA-TARE loci are highly similar; the average pairwise identity between sequences is 88.1% [**Supplementary Figure 2.4**].

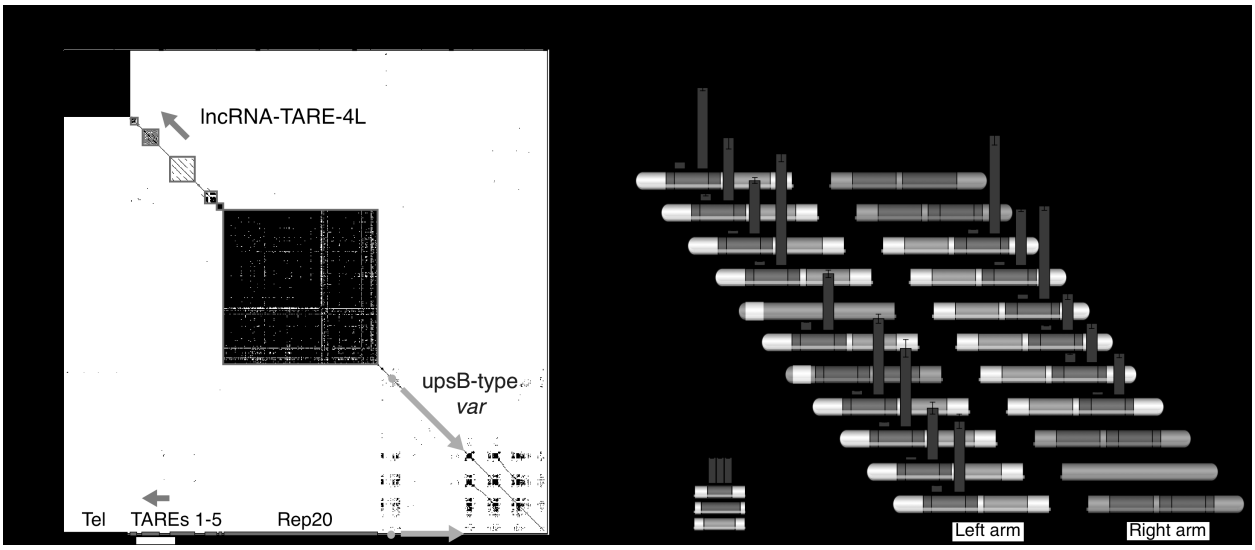


Figure 2.3 | The telomere-associated lncRNA-TARE family neighbors upsB-type *var* loci and is coordinately expressed during parasite schizogony. (a) DNA dot plot analysis of the first 50,000 bp on chromosome four pinpoints lncRNA-TARE-4L in the (telomere-associated repetitive element) TARE 2-3 region between the telomere (Tel) and a silenced upsB-type *var* virulence gene. **(b)** Subtelomeric features are diagrammed on the eleven chromosomes from which we were able to specifically amplify lncRNA-TARE sequence. Features are not to scale, and grayed out chromosome ends were not probed. Notably, an upsB-type *var* gene (orange) is located adjacent to each predicted lncRNA-TARE gene (red). Furthermore, qRT-PCR amplification of fifteen distinct lncRNA-TARE genes shows a coordinated pattern of late stage expression. LncRNA-TARE relative expression is plotted in time-points T20, T30, and T40 +/ 3 hpi with respect to T30. Error bars represent the propagated standard error of the mean (pSEM) from two biological replicate experiments. Housekeeping gene PF08_0085 was used as the reference gene.

LncRNA-TARE Loci are Coordinately Expressed from at least Fifteen Chromosome Ends

Given our evidence for expression of the TARE 2-3 region on chromosome four and conservation of this region on twenty-two *P. falciparum* chromosome ends, we hypothesized that the twenty-two TARE 2-3 regions may be coordinately expressed. We thus set out to further investigate the expression properties of lncRNA-TARE loci in independent blood stage time courses. In line with our array results, we expected lncRNA-TARE-4L, and potentially other lncRNA-TARE genes, to be differentially expressed after parasite DNA replication.

We conducted two additional highly synchronous *P. falciparum* time courses focused on stage-specific time-points T20 (ring), T30 (trophozoite), and T36/T38/T40 (schizont) +/-3 hpi, and subsequently isolated RNA from each stage. We were able to design specific primer pairs targeting fifteen chromosome ends and used qRT-PCR to probe expression at the TARE 2-3 region. Primer pairs were excluded if they did not have at least 90% amplification efficiency or amplified non-specific products [**Methods 2.5** and **Supplementary Table 2.5**].

We found that the TARE 2-3 region is expressed on all fifteen distinct chromosome ends tested. Remarkably, all fifteen lncRNA-TARE genes are coordinately and strongly induced after parasite DNA replication, with maximal lncRNA-TARE transcript abundance observed in the T40 +/- 3 hpi time-point [**Figure 2.3b**]. This result is the first quantitative experimental evidence showing that fifteen, if not more, *P. falciparum* chromosome ends are transcriptionally active between the parasite's DNA replication and cell division cycles. Taken together, we have validated and expanded the microarray discovery of lncRNA-TARE-4L to define a novel telomere-associated lncRNA family, termed lncRNA-TARE. We have also shown that lncRNA-TARE is maximally expressed during an important stage of parasite blood stage development.

RLM-RACE Defines Two Dominant Long Telomere-Associated Transcripts

We next pursued RLM-RACE to map putative transcriptional start and termination sites of lncRNA-TARE. We used a priming strategy in which both the 5' and 3' RACE reactions were primed using the same primer sequence, albeit reverse complemented, to ensure amplification of a contiguous long transcript. Additionally, as we aimed to investigate the transcript boundaries of multiple lncRNA-TARE loci in parallel, we designed RACE primers to simultaneously target twenty lncRNA-TARE sequences [**Figure 2.4**, black triangles, and **Supplementary Table 2.5**].

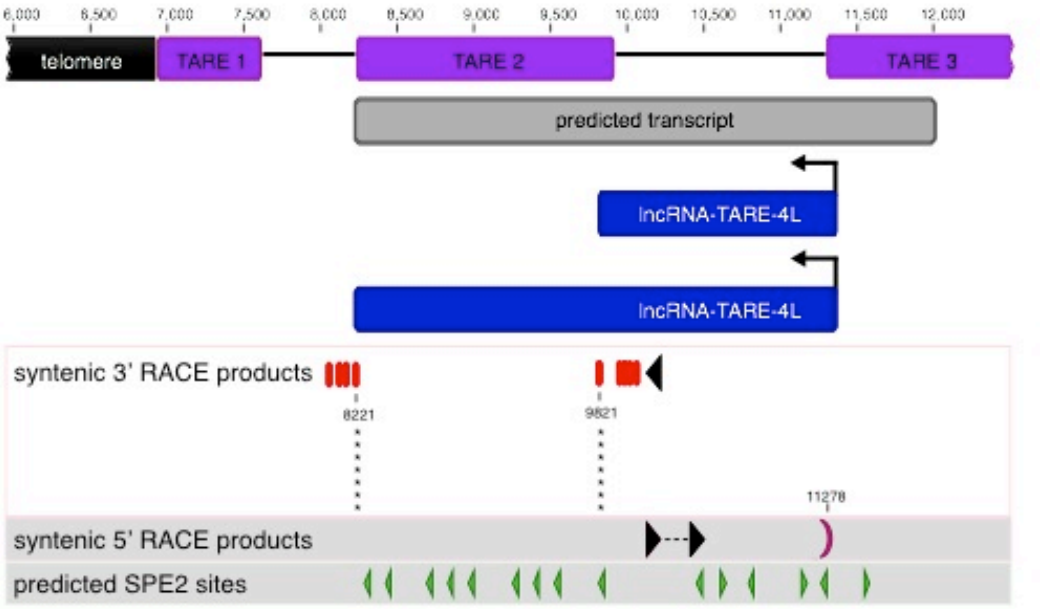


Figure 2.4. RLM-RACE mapping suggests two long telomere-associated transcripts are transcribed in the TARE 2-3 region. 3' RACE products of seven distinct lncRNA-TARE sequences terminated at a precise syntenic site on either side of TARE 2 (stars). Taken together with the 5' RACE product suggests two species approximately 1.5 kb and 3.1 kb in length are transcribed towards the telomeres from the TARE 3 boundary. Genome-wide, the palindromic SPE2 DNA motif (green triangles) is enriched only in the TARE 2-3 region from which lncRNA-TARE is transcribed and in the upstream promoter element of upsB-type *var* genes. Primers are depicted as black triangles.

3' RACE analysis suggested two dominant lncRNA-TARE species are transcribed from the centromere towards the telomere on at least twelve different chromosome ends [**Figure 2.4**, **Supplementary Table 2.6**, and **Supplementary Figure 2.5**]. Interestingly, we found that the longer transcript corresponded to termination immediately after the TARE 2 element, while the shorter transcript corresponded to termination immediately prior to TARE 2. Additionally, the syntenic termination site for seven distinct lncRNA-TARE family members encompassing TARE 2 was identical and corresponded to the transcript model predicted by our sliding-window algorithm [**Methods 2.5**]. In terms of the shorter transcript, we again found that seven distinct lncRNA-TARE species shared a precise syntenic termination site just upstream of TARE 2 [**Figure 2.4**, stars].

We discovered a single putative 5' transcriptional start site for the lncRNA-TARE locus on the left end of chromosome three. Notably, this start site corresponded to the boundary of the TARE 3 element [**Figure 2.4**, **Supplementary Table 2.6**, and **Supplementary Figure 2.5**]. The intrinsic tendency of PCR amplification of multi-gene families to result in strongly biased representation of certain family members [44] likely explains the homogeneity of our cloned 5' RACE products.

Collectively, our RLM-RACE results provide strong evidence that long non-coding RNA genes are present in the *P. falciparum* TARE 2-3 subtelomeric repeats. Furthermore, our results support a model of unidirectional transcription towards the telomeres, as we were unable to amplify transcripts of the opposite polarity. We note, however, that our results do not explicitly rule out bidirectional transcription or the presence of alternative transcript models beyond the approximately 1.5 kb and 3.1 kb species defined here. For example, the high AT content of the *P. falciparum* genome poses a technical barrier for 3' RACE analysis, as internal A-rich regions may hybridize with the 3' RACE primer targeted to poly-A transcript tails.

LncRNA-TARE Loci are Enriched with Transcription Factor Binding Sites

An additional salient feature of lncRNA-TARE is that the TARE 2-3 subtelomeric region contains approximately fifteen occurrences of the bipartite palindromic SPE2 motif on average [45]. This pattern seems to be non-random as there are only two SPE2-enriched loci along the *P. falciparum* genome: the TARE 2-3 region and the upstream promoter element of upsB-type *var* genes. Notably, these two loci account for 94% of 777 predicted SPE2 consensus sites, and both the presence and position of SPE2 sites is conserved on intact chromosome ends [**Figure 2.4** and **Supplementary Figure 2.4**] [11,43,45,46].

Interestingly, Flueck et al. and others have recently demonstrated that a member of *P. falciparum*'s ApiAP2 transcription factor family specifically binds subtelomeric SPE2 sites in late stage parasites [11,45,47,48]. We thus further investigated, using qRT-PCR, the expression of lncRNA-TARE-4L and the *P. falciparum* SPE2-Interacting Protein PfSip2 (PFF0200c). We compared expression at T30 +/- 3 hpi during peak DNA replication to expression at T36, T38, and T40 +/- 3 hpi during parasite schizogony. We found lncRNA-TARE-4L and PfSip2 to have highly correlated late stage temporal profiles. Our results also suggested that PfSip2 is expressed prior to maximal lncRNA-TARE-4L expression, which may indicate PfSip2 induction of the lncRNA-TARE locus and/or co-activation of lncRNA-TARE [**Supplementary Figure 2.6**].

2.3 Discussion

In the present work we have identified and characterized several previously undetected lncRNAs in *P. falciparum,* such as a novel family of twenty-two homologous lncRNA-TARE genes that exhibit coordinated expression at a key stage in the *P. falciparum* life cycle. This family of lncRNA-TAREs encompasses the majority of known binding sites (SPE2) for the ApiAP2 transcription factor PfSip2 [**Figure 2.4** and **Supplementary Figure 2.4**]. As PfSip2 is expressed and specifically binds subtelomeric SPE2 sites at the stage of maximal lncRNA-TARE transcription, these results suggest that PfSip2 binding may positively regulate the lncRNA-TARE locus [45]. Alternatively, lncRNA-TARE transcription may allow PfSip2 binding, or in the case that both models are valid, a regulatory feedback loop between PfSip2 binding and lncRNA-TARE transcription may exist [**Figure 2.5** ***model one***] [49].

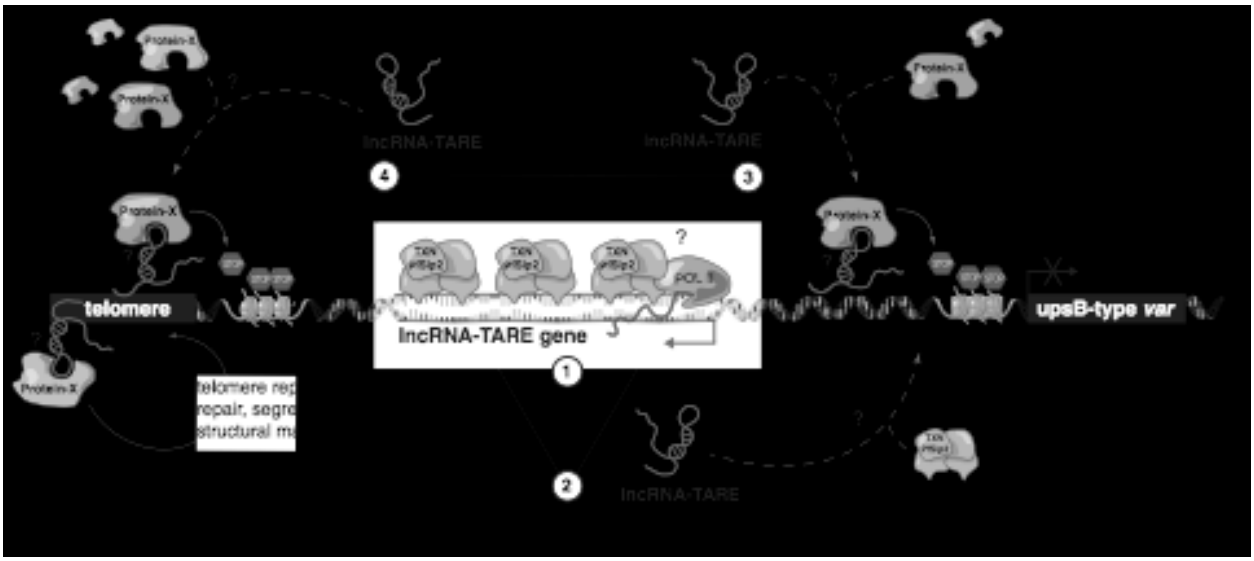


Figure 2.5 | Potential models for lncRNA-TARE at *P. falciparum* chromosome ends. (1) PfSip2 binding at SPE2 sites induces lncRNA-TARE transcription and/or vice-versa. **(2)** LncRNA-TARE guides multiple copies of PfSip2 to chromosome ends. **(3)** LncRNA-TARE recruits chromatin-remodeling factors to initiate post-S-phase memory marks at upsB-type *var* loci. **(4)** LncRNA-TARE facilitates telomeric heterochromatin assembly and/or interacts with telomere-associated proteins.

The only other conserved cluster of SPE2 sites is in the upstream promoter element of upsB-type *var* genes. This suggests that lncRNA-TARE may play a role in subtelomeric upsB-type *var* gene regulation. Consistent with this notion, all twenty-two lncRNA-TARE genes are encoded adjacent to an upsB-type *var* gene. Moreover, PfSip2 is implicated in silenced subtelomeric *var* gene regulation [45]. Thus, lncRNA-TARE may help regulate upsB-type *var* genes by directly or indirectly interacting with and/or recruiting multiple copies of PfSip2 to chromosome ends [**Figure 2.5 *model 2***] [50].

Mutually exclusive *var* gene regulation is known to involve nuclear repositioning and heritable chromatin memory marks [7,51-56]. An emerging theme in mammalian epigenetic regulation is the association of lncRNAs with chromatin modifying complexes, which in turn recruits these factors to their sites of action [37,57-59]. Notably, the induction of lncRNA-TARE transcription between *P. falciparum* DNA replication and cell division is ideally timed to help

initiate post-S-phase epigenetic memory marks at nearby upsB-type *var* genes. Thus, it is possible that lncRNA-TARE may interface with chromatin reading and/or writing factors to modulate the precise epigenetic regulation of nearby subtelomeric *var* loci [**Figure 2.5 *model 3***].

Non-coding transcripts expressed from a conserved bidirectional promoter in *var* gene introns provide some precedent for the proposed link between lncRNA-TARE and *var* gene regulation. Perhaps non-coincidentally, both lncRNA-TARE and *var* gene 'sterile' transcripts have a similar post-S-phase transcriptional profile [60]. However, while the twenty-two lncRNA-TARE loci exclusively pair with upsB-type *var* genes, the previously described *var* gene 'sterile' transcripts are expressed from both silenced and activated *var* genes of all types [39,61]. Thus, lncRNA-TARE may provide a different and/or additional level of regulation unique to subtelomeric upsB-type *var* loci. We also note that in a complementary study by Raabe et al., small subtelomeric ncRNAs ranging in size from 31-103 nucleotides were identified in this region [27]. Taken together, our results perhaps support a model of long precursor transcripts being processed into small ncRNAs.

Interestingly, the origin and transcriptional profile of lncRNA-TARE is strongly reminiscent of the eukaryotic non-coding RNA family TERRA. Across numerous eukaryotic model systems, TERRA is transcribed from subtelomeric loci, associates with telomeric repeats, and is implicated in telomere structural integrity and heterochromatin assembly [21-26,62-67]. Consistent with a possible TERRA-like role for lncRNA-TARE, large chromatin state changes occur at *P. falciparum* chromosome ends in concordance with maximal lncRNA-TARE transcription [**Figure 2.5 *model* 4**] [68,69]. Moreover, TERRA transcript levels are regulated by the cell cycle, with lowest TERRA abundance observed during DNA replication [23,70]. We

similarly observe the lowest lncRNA-TARE transcript levels at T30 +/- 3 hpi, which corresponds to peak parasite DNA replication.

In summary, several pieces of evidence point to a key role for lncRNA-TARE in transcriptional and/or epigenetic regulation of *P. falciparum* telomeric and subtelomeric regions. This work motivates further experimentation to resolve the mechanistic details of lncRNA-TARE and other promising *P. falciparum* lncRNA candidates.

2. 4 Conclusion

Our global transcriptional profiling study provides a catalyst for in-depth functional analyses of high-confidence P. falciparum lncRNA candidates and for a full-genome investigation of parasite strains beyond 3D7. We have identified and characterized 60 putative lncRNAs using conservative thresholds for statistical and biological significance, providing 23 strong candidates for further functional experimentation such as RNA binding assays and expression knockdown experiments. Such next studies will be key in establishing a direct molecular link between specific lncRNA transcripts and parasite proteins and in determining genes modulated by lncRNAs during P. falciparum blood stage development. Recent studies in model eukaryotes suggest that lncRNAs represent a new paradigm in genome regulation and chromatin remodeling. Hence, profiling the non-coding transcriptome of drug-resistant parasites, parasites with mis-regulated virulence gene phenotypes, and hyper-virulent clinical isolates is an exciting new research direction in the quest to eradicate malarial disease.

2.5 Methods

Microarray Design

The DNA tiling array was designed in conjunction with Roche NimbleGen to tile a portion of the P. falciparum genome with a targeted median probe spacing of 12 bp. The 3D7 assembly (PlasmoDB v5.5) was used as the reference sequence. The design targeted all of chromosomes 2, 4, 7 and 9 and partial chromosomes 3:106138-147339, 5:947885-end, and 12:start-66805 [**Supplementary Figure 2.1**]. The three chromosomes tiled in their entirety were selected at random. However, the four partially tiled chromosomes represent regions of particular interest given our hypothesis that lncRNAs may be involved in chromatin remodeling and clinically important parasite processes. Probes were variable length T_m-matched long oligonucleotides, averaging 55 bp each. Probe sequences were screened for excessive cross hybridization to 3D7 sequence: any probes with more than five close SSAHA (Sequence Search and Alignment by Hashing Algorithm) matches were eliminated [71,72]. The final design filled 366479 probes on the array, 96.81% of which are unique. The raw and normalized data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE27937 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27937).

Parasite Culture, RNA Preparation, and cDNA Labeling for Microarray Hybridizations

A clone of P. falciparum strain 3D7 was cultured using standard methods [73,74] and total RNA isolated from each sample as described [75]. Total RNA was cleaned up with an RNeasy column (Qiagen) and concentrated in a Microcon YM-30 centrifugal filter (Millipore). 1 μg of total RNA was then subjected to poly-A selective amplification using Message Amp II

(Ambion), substituting biased dNTP/NTP mixes (2A/T/U:1C/G) for the solutions provided. The resulting aRNA was labeled with Superscript II reverse transcriptase (Invitrogen), using random hexamers, and either Cy3- or Cy5-dUTPs (Amersham) for 2 hours at 42°C following a 10 minute primer annealing step at 65°C. The reaction was concentrated on a Microcon YM-30 column and subjected to array hybridization per standard NimbleGen protocol.

Data Normalization and Quality Control

Raw data from each sample was quantile normalized [76] and log2 transformed prior to prediction of transcriptionally active regions (TARs) along the P. falciparum genome. We based our pre-processing pipeline on established quality control metrics: removal of non-biological variation and a strong correlation between raw and normalized data [77]. **Supplementary Figure 2.7** shows log2(intensity) distributions of each sample before and after quantile normalization, log2(intensity) boxplots of each sample after quantile normalization, and all pair-wise correlation scatterplots of data before and after normalization. Pearson correlation is equal to 1.0 between raw and normalized data in all matched samples. **Supplementary Figure 2.8** shows that for each sample we observed only a minimal increase in median intergenic probe hybridization intensity with number of GC bases. Given that we make no absolute or quantitative expression comparisons of transcripts (only relative expression comparisons of the same transcript across time-points) and we confirm no GC content bias in predicted lncRNAs, we deemed this inconsequential. Normalized data was median centered (at zero) prior to expression profiling and data browsing in Integrated Genomics Viewer (IGV) [78].

Detection of Transcriptionally Active Regions (TARs) from Tiling Arrays

We wrote and implemented an iterative sliding window algorithm to scan each sample's normalized probe hybridization intensity values for statistically significant TARs. Specifically,

we used a single-step maxT permutation procedure (1000 permutations) to transform the mean probe intensity score 'T' calculated in each of the approximately 366k possible window slides along normalized data into a multiple-hypothesis adjusted p-value [30-32]. We then discarded windows with adjusted p-values greater than .05 to control the family-wise error-rate (FWER) of windows predicted to be significant at 5%. We repeated this procedure a total of twenty-eight times using window sizes of 5, 10, 15, 20, 25, 30, and 40 probes. Next, we intersected all significant windows with PlasmoDB v6.5 gene annotations, and merged overlapping annotated windows to define the boundaries of 1229 protein-coding TARs and 8 ribosomal RNA TARs. Similarly, we merged overlapping windows that did not overlap any known or predicted gene to define the boundaries of 123 un-annotated TARs. BEDTools v4 was used for all data intersections and unions [79].

Filtering Un-Annotated TARs

We filtered un-annotated TARs by setting a minimum length criterion of 200 base pairs and ensuring no BLASTX predicted coding potential. Out of 123 predicted un-annotated TARs, 46 were under 200 base pairs in length [**Supplementary Table 2.2**]. We retrieved FASTA sequence for the remaining 77 from plasmodb.org (v6.5), and used the NCBI BLASTX web server to search for any significant protein matches [80]. Default BLASTX settings (BLOSUM62, word size 3, low complexity filtering, etc.) were used except the Expect threshold for reporting match significance (i.e. coding potential) was set at .01. Seventeen sequences with an Expect score < .01 were categorized as putative novel P. falciparum genes or pseudogenes and were excluded from further lncRNA characterization [**Supplementary Table 2.3**]. We searched both the Swissprot and Non-redundant protein sequence (nr) databases with the

following organism queries: all organisms, *Plasmodium, Plasmodium falciparum,* and *Plasmodium falciparum* strain 3D7.

Gene Ontology (GO) Term Analysis of Stage-Specific Genes

GO term analysis of stage-specific genes was performed using GOstat with default settings [81] and Sanger GeneDB P. falciparum gene annotations. Stage-specific genes were determined by intersection of PlasmoDB v6.5 gene annotations with protein-coding transcripts maximally expressed in each time-point. We looked for overrepresented GO terms in stage-specific genes versus the 1360 protein-coding genes covered by the array. **Supplementary Table 2.1** lists the top four most overrepresented GO terms in each time-point and respective p-values.

Evolutionary Sequence Conservation

FASTA sequence for the 60 putative lncRNAs and 8 ribosomal RNA transcripts was retrieved from the 3D7 reference sequence (PlasmoDB v7.1). We also downloaded genomic FASTA sequence from PlasmoDB v7.1 representing all eight sequenced or partially sequenced Plasmodium species (*P. falciparum, P. reichenowi, P. gallinaceum, P. knowlesi, P. vivax, P. berghei, P. yoelli,* and *P. chaubadi*). We searched for sequence conservation using BLASTN (WU-BLAST 2.0MP-WashU [04-May-2006]) using the same low-complexity filtering and context parameters as the BLAST server on plasmodb.org (-filter seg -ctxfactor 2.00) and setting the Expect threshold for significance to 0.01. We recorded the lowest BLASTN p-value within each species [**Supplementary Table 2.4**].

The broad conservation of lncRNA-TARE-4L across all eight Plasmodium species was determined to be significant by null permutations. 600 random intergenic regions were chosen

from the 3D7 reference genome (based on v7.1 annotation). These intergenic regions were sized to match the length distribution of the 60 putative lncRNAs, and were included in the WU-BLAST search. Out of 600 random intergenic regions, we found only 27 to be conserved across all eight species, yielding an empirical p-value of 0.045.

Expression Profiling

To profile each predicted TAR, we calculated its expression in each time-point as the mean hybridization intensity of probes tiling within or up to 25 base pairs on either side of the predicted TAR start and stop coordinates. The expression profile of each TAR was then mean centered across time-points and visualized using a non-hierarchical clustering dimension reduction algorithm. Specifically, we used non-Metric Multi-Dimensional Scaling (nMDS) as implemented in the R-project 'NeatMap' package to order rows and preserve data topology. In development and validation of the 'NeatMap' package, Rajaram et al. have similarly applied nMDS to visualize yeast cell cycle expression data [82]. We point the reader to Figure 1d of Rajaram et al. [82] and Figure 4 of Taguchi et al. [83] for examples and thorough discussion of the utility of nMDS in determining relational patterns of gene expression. Notably, because nMDS is a non-linear numerical optimization technique, multiple ordinations were run to select the optimal solution.

We conducted a detailed comparison of lncRNA and protein-coding expression, finding lncRNA candidates to be expressed on par with protein-coding transcripts. Included in **Supplementary Figure 2.2** are additional visualizations of protein-coding transcript versus putative lncRNA expression, including standard heatmaps and nMDS ordinated heatmaps without mean centering across time-points. **Supplementary Figure 2.2** also provides a histogram of the maximum expression values for lncRNA candidates and protein-coding

transcripts. Notably, we found 30 lncRNA candidates (50%) to be induced by greater than 2-fold across our time course samples [**Supplementary Table 2.4**]. By comparison, 309 of 1229 protein-coding transcripts (25%) match this criterion.

Nearest-Neighboring Genes

Nearest neighboring genes to the set of 60 putative lncRNAs were extracted using the Cistrome Analysis Pipeline (AP) Module and PlasmoDB v6.5 gene coordinates.

Correlation Analysis

To infer putative lncRNA splicing or UTR relationships with neighboring coding genes, we measured the Pearson correlation between putative lncRNA and neighboring coding gene expression profiles [**Supplementary Table 2.4**]. We conservatively defined the neighboring coding gene to be the highest correlated, expressed gene to either side of each putative lncRNA locus. We also examined a null distribution of correlations from adjacent pairs of coding genes. We found the 60 candidate lncRNAs to be enriched for high correlation to neighboring genes. 40 of these candidates were highly correlated ($r > 0.9$), whereas only ten should be highly correlated as demonstrated by our null distribution [**Supplementary Figure 2.3**]. We then further investigated the expression profiles of lncRNA candidates with $r < 0.9$ to ensure correlation values reflected biologically meaningful variation. We defined biologically meaningful variation as a greater than 0.5-fold change across time-points.

Mapping of Homologous lncRNA-TARE Sequences

We used the PlasmoDB v6.5 BLASTN web server to record coordinates for homologous lncRNA-TARE sequences based on the predicted lncRNA-TARE-4L sequence. We then retrieved FASTA sequence of the most telomere-proximal 50000 bp on each chromosome end

from plasmodb.org (v6.5), and used JDotter [84] software to create DNA dotplots mapping the telomeric repeats, TAREs 1-5, Rep20, and first predicted gene on each end. We placed each predicted lncRNA-TARE gene onto the dotplot maps to confirm that lncRNA-TARE maps to TARE 2 and the sequence between TARE 2 and TARE 3 on twenty-two chromosome ends. We then used Geneious to cluster (ClustalW) and investigate the conservation of lncRNA-TARE sequences.

Parasite Culture for qRT-PCR Analysis

Two independent biological replicate time courses were performed to validate and investigate lncRNA-TARE expression in more detail. For each time course, a freshly thawed P. falciparum strain 3D7 clone was cultured using standard methods [73] in human red blood cells (RBCs) at 4% hematocrit. RPMI-HEPES medium was supplemented with 5% human serum (O+) and 5% Albumax II (Gibco). Cultures were initially synchronized using two 5% sorbitol solution treatments [74] spaced by sixteen hours. To then obtain highly synchronized cultures, newly formed ring-stage parasites were selected for using 5% sorbitol solution treatments during the subsequent two re-invasion generations. Highly synchronized cultures were expanded and harvested at stage-specific time-points. Each harvested culture was centrifuged at 2400 rpm in a Sorvall RT6000B, and packed RBCs (2 ml) lysed using a 0.05% saponin solution. Liberated parasites were washed using phosphate-buffered saline (pH 7.4), pelleted at 13.2 RPM in a microcentrifuge, resuspended in 1 ml TRIZOL reagent, and stored at -80°C prior to RNA extraction.

RNA Preparation for qRT-PCR and RLM-RACE Analysis

TRIZOL-chloroform extraction was performed and the aqueous layer applied to an RNeasy column (Qiagen). On-column DNAse digestion was carried out for thirty minutes to

remove genomic DNA. Eluted RNA was also treated with TURBO DNase (Ambion) and cleaned up on a second RNeasy column (Qiagen) to yield high-purity RNA samples.

Quantitative Real-Time PCR (qRT-PCR) Analysis

1 μg of RNA from each time course sample was reverse transcribed using a random priming strategy (Applied Biosystems cDNA High Capacity Reverse Transcription kit) along with a minus RT control reaction for each sample to confirm genomic DNA removal. qPCR reactions were carried out using 800 nM of primers and Roche FastStart SYBR Green Master (Rox) mix. Primer annealing and extension (55°C/60 seconds) was carried out for 40 cycles on an Applied Biosystems 7900 instrument.

We used PCR Miner software [85] to calculate both the cycle threshold (Ct) of each qPCR reaction and the amplification efficiency of each primer pair. We then calculated the relative expression of each lncRNA-TARE gene in each time course sample by averaging technical replicates and using the reference gene PF08_0085 and reference time-point T30 (trophozoite) for normalization. The error of normalized expression ratios was calculated using the delta method, based on a truncated Taylor series expansion, to account for technical variability in both the target and reference gene measurements. Biological replicate experiments were analyzed in isolation and then normalized expression measurements were averaged. We used a Taylor limited expansion method to determine how error propagated in the average expression value.

qRT-PCR Primer Design

Primer pairs to amplify predicted lncRNA-TARE genes and the SPE2-binding protein PfSip2 (PFF0200c) were designed using Premier Biosoft International AlleleID 7.6 software.

AlleleID primer design software carries out highly specific primer design by BLAST searching sequences and masking redundant regions prior to primer design. We also independently verified primer specificity using BLASTN on the plasmodb.org (v6.5) website, and ensured single amplicon melting curves and no primer dimer formation. We required primer pair amplification efficiency, as calculated by PCR Miner software [85], to be at least 90% to ensure reproducible results. We used the previously described housekeeping gene P08_0085 (ubiquitin conjugating enzyme 1) [75] to calculate all normalized relative gene expression ratios. LncRNA-TARE and PfSip2 primer sequences are listed in **Supplementary Table 2.5**.

RLM-RACE

We employed RNA Ligase Mediated Rapid Amplification of cDNA Ends (RLM-RACE) following manufacturer specifications (Ambion) and using 10 μg of T40 +/- 3 hpi RNA mixed 1:1 from two independent time course extractions. We used Premier Biosoft International AlleleID 7.6 software to design primers targeting twenty lncRNA-TARE loci [**Supplementary Table 2.5**]. To map the putative 5' cap, we used a nested priming strategy with primers spaced roughly 350 bp antisense to the target sequence. To map 3' termini, we used a semi-nested priming strategy using a single antisense primer to the target sequence and nested primers corresponding to the 3' adapter sequence. Notably, the 5' RACE outer primer is the reverse complement of the 3' RACE primer, ensuring capture of contiguous transcripts. Minus RT control reactions were included for 3' RACE.

Outer and inner 5' RACE PCR cycling was performed using SuperTaq Plus polymerase (Ambion) and the following cycling conditions: 94°C for 3 min, 5 cycles of 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 3 minutes, 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds, 68°C for 3 minutes, and a final extension for 10 minutes at 68°C. 3' RACE PCR cycling was

analagous except denaturation was performed at 94°C for 15 seconds and extension was performed at 68°C for 8 minutes. PCR products were gel excised, purified using QIAGEN MinElute Gel Extraction Cleanup columns, and cloned into the pCR-2.1TOPO vector (Invitrogen).

27 and 10 colonies corresponding to 3' and 5' RACE products, respectively, were sequenced using Genewiz services and analyzed using Geneious software [86]. A total of 12 different lncRNA-TARE loci were unambiguously represented in sequenced 3' RACE products. The original chromosome and syntenic terminus coordinates on the left end of chromosome four for each sequenced RACE product is included in **Supplementary Table 2.6** along with a graphical alignment of each sequenced RACE product to the left end of chromosome four in **Supplementary Figure 2.5**. RACE products were trimmed to exclude any low-quality base calls and vector sequence beyond the first four bases prior to alignment.

2.6 References

1. Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, et al. (2011) A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. Genome biology 12: R56.
2. Coleman BI, Duraisingh MT (2008) Transcriptional control and gene silencing in Plasmodium falciparum. Cellular Microbiology 10: 1935-1946.
3. Horrocks P, Wong E, Russell K, Emes RD (2009) Control of gene expression in Plasmodium falciparum - Ten years on. Molecular and Biochemical Parasitology 164: 9-25.
4. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu JC, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. Plos Biology 1: 85-100.
5. Le Roch KG, Zhou YY, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503-1508.
6. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, et al. (2010) New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Molecular Microbiology 76: 12-24.
7. Scherf A, Lopez-Rubio JJ, Riviere L (2008) Antigenic variation in Plasmodium falciparum. Annu Rev Microbiol 62: 445-470.
8. Baum J, Papenfuss AT, Mair GR, Janse CJ, Vlachou D, et al. (2009) Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. Nucleic Acids Res 37: 3788-3798.
9. Cui L, Miao J (2010) Chromatin-mediated Epigenetic Regulation in the Malaria Parasite Plasmodium falciparum. Eukaryot Cell.
10. Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res 33: 3994-4006.
11. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M (2010) Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. PLoS Pathog 6: e1001165.
12. Chaal BK, Gupta AP, Wastuwidyaningtyas BD, Luah YH, Bozdech Z (2010) Histone Deacetylases Play a Major Role in the Transcriptional Regulation of the Plasmodium falciparum Life Cycle. Plos Pathogens 6: -.
13. Merrick CJ, Duraisingh MT (2010) Epigenetics in Plasmodium: what do we really know? Eukaryot Cell 9: 1150-1158.
14. Coulson RM, Hall N, Ouzounis CA (2004) Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. Genome Res 14: 1548-1554.
15. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.
16. Corcoran AE (2010) The epigenetic role of non-coding RNA transcription and nuclear organization in immunoglobulin repertoire generation. Semin Immunol 22: 353-361.
17. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464: 1071-1076.
18. Pandey RR, Mondal T, Mohammad F, Enroth S, Redrup L, et al. (2008) Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation. Molecular Cell 32: 232-246.
19. Brockdorff N, Ashworth A, Kay GF, Mccabe VM, Norris DP, et al. (1992) The Product of the Mouse Xist Gene Is a 15 Kb Inactive X-Specific Transcript Containing No Conserved Orf and Located in the Nucleus. Cell 71: 515-526.
20. Lee JT (2009) Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. Genes Dev 23: 1831-1842.
21. Schoeftner S, Blasco MA (2008) Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol 10: 228-236.
22. Schoeftner S, Blasco MA (2010) Chromatin regulation and non-coding RNAs at mammalian telomeres. Semin Cell Dev Biol 21: 186-193.
23. Feuerhahn S, Iglesias N, Panza A, Porro A, Lingner J (2010) TERRA biogenesis, turnover and implications for function. FEBS Lett 584: 3812-3818.

24. Redon S, Reichenbach P, Lingner J (2010) The non-coding RNA TERRA is a natural ligand and direct inhibitor of human telomerase. Nucleic Acids Res 38: 5797-5806.
25. Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. Science 318: 798-801.
26. Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM (2009) TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. Mol Cell 35: 403-413.
27. Raabe CA, Sanchez CP, Randau G, Robeck T, Skryabin BV, et al. (2010) A global view of the nonprotein-coding transcriptome in Plasmodium falciparum. Nucleic Acids Research 38: 608-617.
28. Goldberg DE, Cowman AF (2010) Moving in and renovating: exporting proteins from Plasmodium into host erythrocytes. Nature Reviews Microbiology 8: 617-621.
29. Maier AG, Cooke BM, Cowman AF, Tilley L (2009) Malaria parasite proteins that remodel the host erythrocyte. Nature Reviews Microbiology 7: 341-354.
30. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223-227.
31. Guttman M, Mies C, Dudycz-Sulicz K, Diskin SJ, Baldwin DA, et al. (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. PLoS Genet 3: e143.
32. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Statistical Science 18: 71-103.
33. Chakrabarti K, Pearson M, Grate L, Sterne-Weiler T, Deans J, et al. (2007) Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. RNA 13: 1923-1939.
34. Upadhyay R, Bawankar P, Malhotra D, Patankar S (2005) A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of Plasmodium falciparum. Mol Biochem Parasitol 144: 149-158.
35. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.
36. Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5: e1000459.
37. Koziol MJ, Rinn JL (2010) RNA traffic control of chromatin complexes. Current Opinion in Genetics & Development 20: 142-148.
38. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46-58.
39. Epp C, Li F, Howitt CA, Chookajorn T, Deitsch KW (2009) Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. RNA 15: 116-127.
40. Deitsch KW, del Pinal A, Wellems TE (1999) Intra-cluster recombination and var transcription switches in the antigenic variation of Plasmodium falciparum. Mol Biochem Parasitol 101: 107-116.
41. Figueiredo L, Scherf A (2005) Plasmodium telomeres and telomerase: the usual actors in an unusual scenario. Chromosome Research 13: 517-524.
42. Scherf A, Figueiredo LM, Freitas-Junior LH (2001) Plasmodium telomeres: a pathogen's perspective. Curr Opin Microbiol 4: 409-414.
43. Lavstsen T, Salanti A, Jensen ATR, Arnot DE, Theander TG (2003) Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malaria Journal 2: -.
44. Esumi S, Kaneko R, Kawamura Y, Yagi T (2006) Split single-cell RT-PCR analysis of Purkinje cells. Nat Protoc 1: 2143-2151.
45. Flueck C, Bartfai R, Niederwieser I, Witmer K, Alako BT, et al. (2010) A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. PLoS Pathog 6: e1000784.
46. Painter HJ, Campbell TL, Llinas M (2010) The Apicomplexan AP2 family: Integral factors regulating Plasmodium development. Mol Biochem Parasitol.
47. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, et al. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. Proceedings of the National Academy of Sciences of the United States of America 105: 8393-8398.
48. Voss TS, Kaestli M, Vogel D, Bopp S, Beck HP (2003) Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters. Molecular Microbiology 48: 1593-1607.
49. Bond AM, Vangompel MJ, Sametsky EA, Clark MF, Savage JC, et al. (2009) Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat Neurosci 12: 1020-1027.

50. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP (2010) Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor. Science Signaling 3: -.
51. Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT, et al. (2005) Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in Plasmodium faiciparum. Cell 121: 13-24.
52. Dzikowski R, Deitsch KW (2009) Genetics of antigenic variation in Plasmodium falciparum. Current Genetics 55: 103-110.
53. Dzikowski R, Li F, Amulic B, Eisberg A, Frank M, et al. (2007) Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. EMBO Rep 8: 959-965.
54. Scherf A, Riviere L, Lopez-Rubio JJ (2008) SnapShot: var gene expression in the malaria parasite. Cell 134: 190-U112.
55. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, et al. (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. Cell 121: 25-36.
56. Voss TS, Tonkin CJ, Marty AJ, Thompson JK, Healer J, et al. (2007) Alterations in local chromatin environment are involved in silencing and activation of subtelomeric var genes in Plasmodium falciparum. Mol Microbiol 66: 139-150.
57. Bernstein E, Allis CD (2005) RNA meets chromatin. Genes & Development 19: 1635-1655.
58. Nagano T, Fraser P (2009) Emerging similarities in epigenetic gene silencing by long noncoding RNAs. Mamm Genome 20: 557-562.
59. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. (2010) Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science 329: 689-693.
60. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, et al. (1995) The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82: 89-100.
61. Ralph SA, Bischoff E, Mattei D, Sismeiro O, Dillies MA, et al. (2005) Transcriptome analysis of antigenic variation in Plasmodium falciparum--var silencing is not dependent on antisense RNA. Genome Biol 6: R93.
62. Deng Z, Campbell AE, Lieberman PM (2010) TERRA, CpG methylation and telomere heterochromatin: lessons from ICF syndrome cells. Cell Cycle 9: 69-74.
63. Horard B, Gilson E (2008) Telomeric RNA enters the game. Nat Cell Biol 10: 113-115.
64. Xu Y, Suzuki Y, Ito K, Komiyama M (2010) Telomeric repeat-containing RNA structure in living cells. Proc Natl Acad Sci U S A 107: 14579-14584.
65. Luke B, Lingner J (2009) TERRA: telomeric repeat-containing RNA. EMBO J 28: 2503-2510.
66. Caslini C, Connelly JA, Serna A, Broccoli D, Hess JL (2009) MLL associates with telomeres and regulates telomeric repeat-containing RNA transcription. Mol Cell Biol 29: 4519-4526.
67. Luke B, Panza A, Redon S, Iglesias N, Li Z, et al. (2008) The Rat1p 5' to 3' exonuclease degrades telomeric repeat-containing RNA and promotes telomere elongation in Saccharomyces cerevisiae. Mol Cell 32: 465-477.
68. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, et al. (2010) Nucleosome landscape and control of transcription in the human malaria parasite. Genome Research 20: 228-238.
69. Westenberger SJ, Cui L, Dharia N, Winzeler E, Cui LW (2009) Genome-wide nucleosome mapping of Plasmodium falciparum reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes. Bmc Genomics 10: -.
70. Porro A, Feuerhahn S, Reichenbach P, Lingner J (2010) Molecular dissection of telomeric repeat-containing RNA biogenesis unveils the presence of distinct and multiple regulatory pathways. Mol Cell Biol 30: 4808-4817.
71. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11: 1725-1729.
72. Morgulis A, Gertz EM, Schaffer AA, Agarwala R (2006) WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22: 134-141.
73. Trager W, Jensen JB (1976) Human Malaria Parasites in Continuous Culture. Science 193: 673-675.
74. Lambros C, Vanderberg JP (1979) Synchronization of Plasmodium falciparum erythrocytic stages in culture. J Parasitol 65: 418-420.
75. Sims JS, Militello KT, Sims PA, Patel VP, Kasper JM, et al. (2009) Patterns of gene-specific and total transcriptional activity during the Plasmodium falciparum intraerythrocytic developmental cycle. Eukaryot Cell 8: 327-338.

76. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.
77. Thomassen GO, Rowe AD, Lagesen K, Lindvall JM, Rognes T (2009) Custom design and analysis of high-density oligonucleotide bacterial tiling microarrays. PLoS One 4: e5943.
78. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24-26.
79. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
80. Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. Nat Genet 3: 266-272.
81. Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20: 1464-1465.
82. Rajaram S, Oono Y (2010) NeatMap--non-clustering heat map alternatives in R. BMC Bioinformatics 11: 45.
83. Taguchi YH, Oono Y (2005) Relational patterns of gene expression via non-metric multidimensional scaling analysis. Bioinformatics 21: 730-740.
84. Brodie R, Roper RL, Upton C (2004) JDotter: a Java interface to multiple dotplots generated by dotter. Bioinformatics 20: 279-281.
85. Zhao S, Fernald RD (2005) Comprehensive algorithm for quantitative real-time polymerase chain reaction. Journal of Computational Biology 12: 1047-1064.
86. Drummond AJ AB, Buxton S, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2011) Geneious v4.8, Available from http://www.geneious.com.

Chapter 3 | **Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies extensive long non-coding RNA with developmental implications, including antisense and circular RNA**

Contributions: Kate M Broadbent[1,2§], Jill C Broadbent[3,4], Ulf Ribacke[5,6], Dyann Wirth[2,5], John L. Rinn*[1,2,7§], Pardis C. Sabeti*[1,2,3,4§]

Abstract

The human malaria parasite *Plasmodium falciparum* has a complex and multi-stage life cycle that requires extensive and precise gene regulation to allow invasion and hijacking of human host cells, transmission through the female *Anopheles* mosquito vector, and escape from the immune systems of both vector and host. To date, the regulatory elements orchestrating these critical parasite processes remain largely unknown. However, there is mounting evidence across a broad range of species that intergenic long non-coding RNA (lncRNA) and antisense RNA can regulate chromatin state and gene expression. To pursue such functional roles for lncRNAs in *P. falciparum*, we performed deep, strand-specific RNA sequencing of fifteen non-polyA-selected blood stage samples, and assembled and characterized the properties of 660 intergenic lncRNAs, 474 antisense RNAs, and 1381 circular RNAs (circRNAs). We further validated the non-canonical splice junctions of six *P. falciparum* circRNAs, an emerging class of non-coding RNA with regulatory potential and unexplored functional significance in *P. falciparum*. Our comprehensive analysis of *P. falciparum* lncRNAs indicates a functional role for these transcripts; *P. falciparum* intergenic lncRNAs and antisense RNAs are developmentally

regulated in a similar periodic fashion to annotated transcripts, and sense-antisense pair expression is significantly anti-correlated. Notable outliers include intergenic lncRNAs that strongly peak in expression during parasite invasion, such as the telomere-associated lncRNA-TARE family, antisense RNAs that drop in expression during parasite invasion, and a highly correlated, multi-exonic, antisense counterpart to *P. falciparum* Gametocyte Developmental Protein 1 *(PfGDV1)*. Taken together, our results characterize over two thousand *P. falciparum* intergenic lncRNA, antisense, and circRNA candidates and highlight promising *P. falciparum* lncRNAs for future investigation.

Authors' contributions

KMB, JLR, and PCS conceived of the study and participated in its design, coordination, and interpretation. KMB planned and harvested the morphology-based sample set. UR planned and harvested the 56-hour sample set. KMB and JCB extracted and validated RNA samples, and developed and validated the library preparation protocol. KMB performed all computational analyses relating to transcriptome assembly, lncRNA characterization, and circRNA prediction. KMB designed and performed circRNA validation experiments. DW contributed to data acquisition. KMB wrote the manuscript. This work was supported by grants from the Gates Foundation and the NIH (1DP2OD006514-01, 1DP20D00667-01). KMB was supported by a National Science Foundation (NSF) Graduate Research Fellowship.

Acknowledgements

We would like to thank Niroshini Senaratne of the Harvard School of Public Health for assistance in parasite culture, as well as Jacquin Niles of the Massachusetts Institute of Technology and members of the Broad Institute-Harvard Malaria Initiative, Rinn lab, and Sabeti lab for helpful feedback throughout.

3. 1 Introduction

Plasmodium falciparum is the most deadly human malaria parasite, notorious for its immense disease burden, ability to persist in individuals for months if not longer, and rapid development of resistance to all currently available treatments [1-5]. The symptomatic characteristics of acute *P. falciparum* malaria infection correspond to cycles of red blood cell (RBC) rupture, as merozoite parasites invade RBCs, asexually replicate into 8-36 new daughter merozoites, egress from the RBCs, and repeat the process every 48 hours [6-9]. This process can be readily modeled in the lab, in contrast to the sexual stage required for transmission, which takes 8-12 days in human RBCs and then an additional 8-15 days in mosquitoes [10,11]. Due to the clinical symptoms associated with the asexual blood stage and the relative ease of obtaining samples, the vast majority of current anti-malarial compounds and research programs target this stage of the parasite life cycle [12]. However, the idea of targeting both the symptomatic and transmissible parasite form is garnering increased public attention, making research on sexual stage commitment and sexual development a priority as well [12-14].

The first *P. falciparum* genome sequence was published in 2002 [15]. Our understanding of malaria biology has advanced considerably since this milestone, largely due to genome-wide studies [16,17]. Early transcriptome studies found that key *P. falciparum* proteins are typically transcribed only once per blood stage, 'just-in-time' for translation and function [18,19]. Subsequently, global ribosome profiling and proteome studies revealed significant post-transcriptional regulation and a unique histone code involving at least 44 histone post-translational modifications and four novel histone variants [20-23]. Additionally, paired transcriptome-epigenome studies found dynamic chromatin remodeling and clonally variant gene

expression (CVGE) patterns during blood stage development [24-27], with independent studies confirming a heritable epigenetic layer to monoallelic expression of the 60-member *P. falciparum* Erythrocyte Membrane Protein 1 (PfEMP1*)*-encoding *var* gene family, and Cytoadherence Linked Asexual Gene 3 (*CLAG3)* loci [28-32].

Remarkably, while it has become increasingly clear over the past decade that the *P. falciparum* genome is tightly regulated, the regulatory elements themselves are still largely unknown [33-36]. For example, it is not mechanistically clear how the parasite transcriptionally silences, activates, or switches Pfemp1-encoding *var* genes to evade the human immune system, or how the parasite switches from asexual to sexual development [37,38]. Few sequence-specific transcription factors have been identified, and *P. falciparum* does not encode microRNAs, microRNA processing machinery, or RNA-induced silencing complex (RISC) components [39-41]. With the absence of many known transcription factors and the canonical RNA interference pathway, master regulatory elements orchestrating immune escape, invasion, transmission, and other critical parasite processes remain to be discovered.

We hypothesized that long non-coding RNAs (lncRNAs) may provide a crucial regulatory layer in *P. falciparum* transcriptional, post-transcriptional, and chromatin state control. Recent survey studies have shown prevalent lncRNA transcription in *P. falciparum* [42-47], and a growing body of evidence supports the dominant regulatory roles of lncRNAs in humans and model organisms [48,49]. For example, it has been shown that lncRNAs coordinate X chromosome inactivation in female mammalian cells, flowering time in plants, and gametogenesis in budding yeast [50-55]. However, while prior work has suggested the transcription of intergenic, antisense, and even circular RNA (circRNA) in *P. falciparum*, lncRNA transcript models have not been defined, non-coding properties have not been

generalized on a broad scale, and only two examples of *P. falciparum* circRNAs have been described previously [42-46].

In this study, we aimed to apply recent advancements in next generation sequencing technology to the transcriptome-wide assembly and characterization of *P. falciparum* intergenic lncRNA and antisense RNA, searching for evidence of circRNA transcription as well. The raw and analyzed data, along with the methodology of this study, are broadly useful to the malaria community and should facilitate both *P. falciparum* lncRNA functional investigation and strand-specific sequencing of additional *P. falciparum* samples.

3.2 Results

Strand-Specific RNA Sequencing of Biological Replicate Blood Stage Time Courses

To investigate *P. falciparum* lncRNA transcription, we harvested fifteen blood stage samples from two biological replicate time courses [**Figure 3.1A**]. The first time course comprised eleven samples that finely map temporal changes during *P. falciparum* blood stage development. We harvested samples over 56 hours, at roughly 4-hour time intervals, from a tightly synchronized *P. falciparum* 3D7 parasite population. As the asexual blood stage is an approximately 48-hour cycle, this time course allowed us to profile gene expression during RBC rupture and parasite invasion. The second time course comprised four samples harvested in synchronous *P. falciparum* 3D7 parasites approximately four hours before and after the ring to trophozoite and trophozoite to schizont morphological stage transitions, which occur during the blood stage at 24 hours post-infection (hpi) and 36 hpi, respectively [6-8].

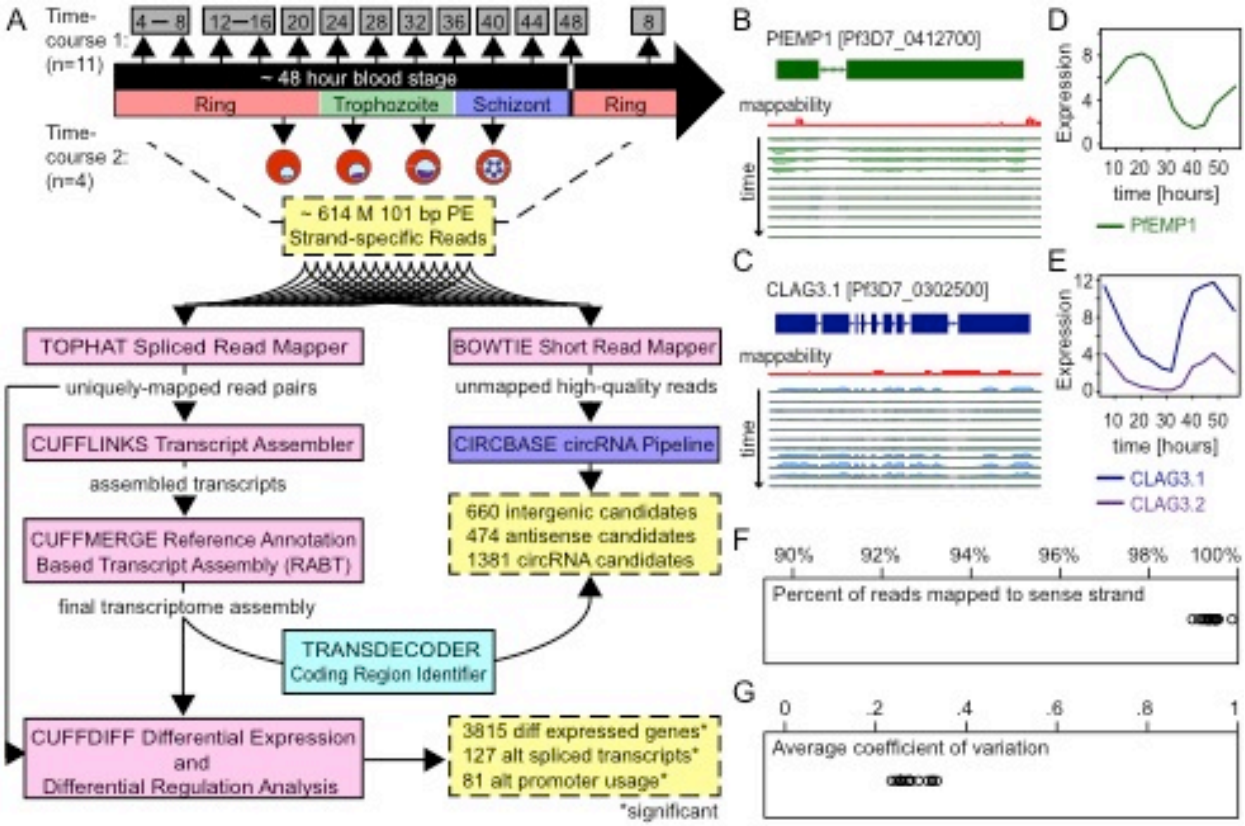


Figure 3.1 | Overview of *P. falciparum* RNA sequencing sample set, computational pipeline, and read alignment metrics. (A) We harvested total RNA from two independent *P. falciparum* blood stage time courses, including a 56-hour time course consisting of eleven samples. We combined samples harvested 4 and 8 hpi at equal ratios (further referred to as T6). Similarly, we combined samples harvested 12 and 16 hpi at equal ratios (further referred to as T14). We harvested four samples from a second time course approximately 4 hours before and after gross stage transitions. Thus these samples correspond to the late ring, early trophozoite, late trophozoite, and early schizont stages, respectively. In total, we sequenced 15 strand-specific RNA sequencing libraries on an Illumina Hiseq 2000 machine. Illumina sequencing yielded approximately 614 million 101-bp paired-end reads. We analyzed reads using the Tuxedo suite (Bowtie, TopHat, Cufflinks, Cuffmerge, and Cuffdiff) and according to the circBase circRNA discovery pipeline. Using this approach, we identified 660 intergenic lncRNA (647 unique loci), 474 antisense RNA (467 unique loci), and 1381 circRNA candidates. Additionally, 3815 genes, 127 transcripts, and 81 promoters reached statistical significance in terms of differential expression, alternative splicing, and alternative promoter usage, respectively. **(B)/(C)** Normalized read alignment tracks across a PfEMP1-encoding *var* gene (*Pf3D7_0212700*) and the *CLAG3.1* gene (*Pf3D7_0302500*) indicated that these challenging loci could generally be (perfectly and uniquely) mapped. Annotated gene models are shown in dark green and dark blue. Reads from each 56-hour time course sample mapping to the (-) strand are shown below each horizontal axis in light green, while reads mapping to the (+) strand are shown above each horizontal axis in light blue. Uniqueness of 100mers is plotted in red as a mappability track, where the baseline represents a score of one, or uniquely mapping.

Figure 3.1 (Continued) | Overview of *P. falciparum* RNA sequencing sample set, computational pipeline, and read alignment metrics. (D)/(E) Plotting the expression during the 56-hour time course of the dominant Pfemp1-encoding *var* gene *(Pf3D7_0212700*) and both the *CLAG3.1* (*Pf3D7_0302500*) and *CLAG3.2* (*Pf3D7*_0502200) genes showed, respectively, that *var* gene expression peaked during the ring stage, whereas *CLAG3.1* and *CLAG3.2* expression peaked during the schizont stage. Moreover, as *CLAG3* genes are mutually exclusively expressed [28,29], we found that that the bulk of our parasites transcribed only the *CLAG3.1* gene. Expression is plotted in units of log2(FPKM+1). **(F)** The percent of reads in each library mapping to annotated transcripts in the proper orientation (per reads mapping to annotated transcripts) ranged from 98.92% to 99.81%. The average calculated from both reads is reported. **(G)** The average coefficient of variation across the top 2000 expressed genes in each library (or roughly top 50% of expressed genes) ranged from 0.23 to 0.33.

Given the high AT content and dense genic structure of the *P. falciparum* genome, we then extensively optimized RNA sequencing procedures, both experimental and computational, in order to derive a high-quality *P. falciparum* transcriptome. In terms of experimental optimization, we tested numerous variables and pursued technical developments shown to reduce sequence-based bias in DNA sequencing libraries and to improve strand-specificity in RNA sequencing libraries [56-62]. Subsequently, we established a library preparation protocol that uses multiple DNase treatments to remove genomic DNA, Ribo-Zero beads to remove ribosomal RNA, the dUTP method with Actinomycin D to preserve strand specificity, and the KAPA HiFi polymerase to amplify libraries in real-time for the minimum number of cycles necessary [see **Methods 3.4**, **Supplementary Figure 3.1**, and **Supplementary Protocol 3.1** for further details].

After harvesting samples from the two time courses and developing and validating our strand-specific library preparation protocol, we prepared libraries from the fifteen blood stage samples in parallel, and sequenced libraries on two lanes of an Illumina Hiseq 2000 machine. Illumina sequencing yielded 614 million 101 base-pair (bp) paired-end reads in total, with sequencing depth ranging from 20 to 30 million (perfectly and uniquely) alignable reads per

sample. We noted high base quality scores and no significant adapter contamination [**Supplementary Table 3.1**, **Supplementary Table 3.2**, and **Supplementary Figure 3.2**].

Aligning and Benchmarking Sequences

We took a conservative approach to read alignment, requiring read pairs to map perfectly and uniquely to the *P. falciparum* 3D7 reference genome. In support of this, we determined that 96.53% of all possible 100mers in the *P. falciparum* genome are unique. In addition, we tested our ability to map read pairs across repeated gene families, such as the Pfemp1-encoding *var* gene family and the two *CLAG3* loci, which we calculated share 96.4% sequence similarity. Specifically, we visualized a lower bound to mappability across these repeated loci by plotting the uniqueness of 100mers as a mappability track. **Figure 3.1B** and **Figure 3.1C** show the mappability track (in red) compared to strand-specific read coverage across a Pfemp1-encoding *var* gene (*Pf3D7_0412700*) and the *CLAG3.1* gene (*Pf3D7_0302500*), respectively. **Figure 3.1C** and **Figure 3.1D** plot the expression profiles of these genes, as well as the *CLAG3.2* gene (*Pf3D7_0302200*). Using this stringent approach and paired-end information, we were able to uniquely map read pairs to these repeated loci, including through short stretches of non-unique sequence.

After conservatively aligning reads using TopHat [63], we assessed data quality following the RNA sequencing benchmarking metrics put forth by DeLuca et al. and Wang et al. [64,65]. We calculated the strand-specificity, coefficient of variation, duplication rate, gap rate, ribosomal RNA rate, exonic rate, insert size, and GC content of each aligned set of reads [**Supplementary Table 3.2**]. Importantly, we found that greater than 98.92% of reads mapped to the reference strand in the expected orientation in each sample [**Figure 3.1F**]. This result was on

par with yeast strand-specific sequencing libraries, and confirmed that our data was highly strand-specific [57]. We also found an average coefficient of variation (CV) of between 0.23 and 0.33 across the top 2000 expressed genes (or roughly top 50% of expressed genes) in each sample [**Figure 3.1G**]. These CV values were lower than the lowest CV value reported in the benchmarking study referenced above (0.54), indicating more even read coverage in our samples [57]. Taken together, the rigorous examination of our data quality demonstrated that it was comparable to the state-of-the-art in model organisms.

Benchmarking Time Courses

Comparing samples between two independent time courses is a known challenge in the field, and can be confounded by experimental factors such as culture conditions [27,66,67]. We thus developed a computational solution that leverages multidimenstional scaling (MDS) to assess stage similarities on a transcriptome-wide scale. While MDS has not previously been used for *P. falciparum* sample comparisons, its utility has been demonstrated in humans and in model organisms such as yeast, especially when periodicity is expected [68,69]. MDS analysis using sample profiles from both time courses revealed the cyclical nature of the *P. falciparum* blood stage, with samples progressing in time around an approximately 48-hour clock [**Figure 3.2A**]. This analysis also confirmed that the four morphology-based samples corresponded to the 56-hour, high-resolution time course samples at expected intervals.

Figure 3.2 | Multidimensional scaling and Gene Ontology confirm expected *P. falciparum* blood stage expression patterns. (A) The MDS plot of sample profiles embedded samples around a circle. Traversing the circle, we found that samples progressed through the approximately 48-hour *P. falciparum* blood stage according to their time and morphology labels as expected. The 56-hour time course samples are labeled in red, green, and blue, with red corresponding to samples harvested within the predicted ring stage, green corresponding to samples harvested within the predicted trophozoite stage, and blue corresponding to samples harvested within the predicted schizont stage. The morphology-based labels correspond to the late ring, early trophozoite, late trophozoite, and early schizont stages, respectively. **(B)** Similarly, the MDS plot of differentially expressed, annotated mRNA gene expression profiles embedded genes around a circle, with their angular position corresponding to their stage of maximum expression. Red indicates 1632 ring-specific genes, green indicates 1378 trophozoite-specific genes, and blue indicates 1274 schizont-specific genes. GO term enrichment analysis on each set of stage-specific genes indicated that genes involved in host cell adhesion processes were expressed during the ring stage, followed by genes involved in DNA replication and metabolic processes during the trophozoite stage, followed by genes involved in DNA replication and protein catabolic processes during the schizont stage. All GO term p-values < .002 **(C)** We visualized the mean-centered expression profiles of annotated mRNA genes during the 56-hour time course in a circular, non-clustering heatmap [68]. Genes are ordered by their angular position in the MDS plot of gene expression profiles and samples are ordered by time. Expression is in units of log2(FPKM+1).

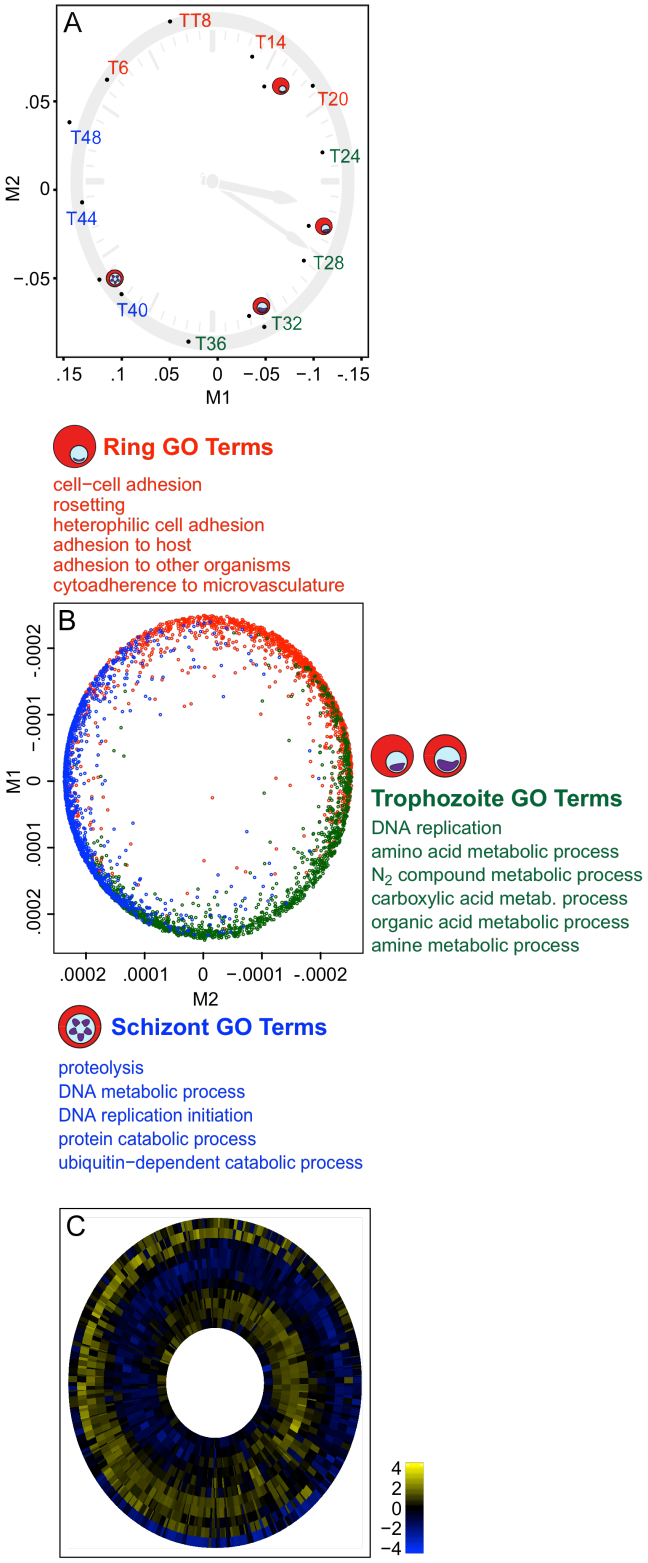


Figure 3.2 (Continued) | Multidimensional scaling and Gene Ontology confirm expected *P. falciparum* blood stage expression patterns.

As a complementary analysis, we performed MDS using differentially expressed annotated messenger RNA (mRNA) gene expression profiles. We classified 1632 ring-specific, 1378 trophozoite-specific, and 1274 schizont-specific genes according to their maximal expression time-point, and color-coded these genes as red, green, and blue, respectively, in the two-dimensional plot produced by MDS [**Figure 3.2B**]. The clear separation of ring-, trophozoite-, and schizont-specific genes indicated that genes were embedded in the MDS plot according to their functional stage of expression. To investigate this further, we ordered genes by their angular position in the MDS plot and samples by time, and visualized gene expression profiles in a circular non-clustering heatmap [**Figure 3.2C**] [68]. This revealed the expected periodic expression cascade of the *P. falciparum* blood stage, with most genes peaking in expression only once.

We next sought to connect the cyclical, periodic expression patterns evident in our MDS results to *P. falciparum* biology. First, we considered the expression profiles of well-established stage-specific marker genes. For example, based on previous studies, the dominant Pfemp1-encoding *var* gene should be maximally expressed during the ring stage, and one of the *CLAG3* genes should be maximally expressed during the schizont stage [29,32]. Our expression profiles matched these expectations [**Figure 3.1F/G**, and **Supplementary Figure 3.3**]. We similarly assessed many additional marker genes, confirming their expected stage-specific expression profiles in each time course [**Supplementary Figure 3.4**]. To gain a more global view of sample biology, we then computed GO term enrichment on the ring-, trophozoite-, and schizont-specific gene sets that we previously defined [70]. Ring-, trophozoite-, and schizont-specific GO terms were specific to host cell adhesion processes, metabolic processes, and protein catabolic processes, respectively, with DNA replication spanning both trophozoite- and schizont-specific

GO terms [**Figure 3.2B**, and **Supplementary Table 3.3**]. These GO terms were highly consistent with our current understanding of *P. falciparum* biology [6-8]. Taken together, MDS paired with gene-by-gene and global GO term enrichment analysis validated the biological integrity of our time course samples.

Transcript Assembly

We next set out to assemble *P. falciparum* transcript structures, with or without the assistance of annotated transcript models, and to assess assembly performance using either Cufflinks or genome-guided Trinity [71-73] [**Figure 3.1A**]. Specifically, we were looking for high contiguity (a high rate of annotated transcripts being spanned by one assembled transcript over at least 90% of the annotated transcript exonic length), low chimerism (a low rate of assembled transcripts spanning more than one annotated transcript), and for the final assembly to be manageable and high-confidence [74,75].

Using Cufflinks without the assistance of annotation, we found that 81.5% of annotated transcripts had assembled transcripts contiguously spanning them, while only 7.9% of assembled transcripts were chimeric [**Table 3.1**]. With the assistance of annotation, the chimerism rate dropped to 6.7% and the contiguity rate naturally rose to 100% [**Table 3.1**]. The number of predicted unannotated transcripts, however, did not appreciably change when we used Cufflinks in reference-annotation based transcript assembly (RABT) mode or without the assistance of annotated transcript models. Based on its performance features, we chose to further explore the high-confidence Cufflinks transcripts (at least 50 supporting reads in at least one sample). However, it may be possible to filter the genome-guided Trinity results based on read support or expression level to yield a more manageable *P. falciparum* transcriptome assembly [**Table 3.1**].

Table 3.1 | Comparative assessment of *P. falciparum* transcriptome assembly highlights the performance of Cufflinks RABT. We compared the contiguity, chimerism, and feature counts of Cufflinks versus Genome-guided Trinity transcriptome assembly, with or without the assistance of annotation. Cufflinks incorporating reference annotation based transcriptome assembly (RABT) provided the optimal *P. falciparum* transcriptome. Contiguity is the rate of annotated transcripts covered by one assembled transcript over at least 90% of the annotated transcript exonic length in the correct orientation. Chimerism is the rate of assembled transcripts that span more than one annotated transcript in the correct orientation. Total number of transcripts, number of intergenic transcripts, and number of antisense transcripts correspond to the total number of assembled transcripts, the number of assembled transcripts predicted between *PlasmoDBv10.0* annotations, and the number of assembled transcripts predicted antisense to *PlasmoDBv10.0* annotations. In total, 5,777 transcripts are annotated in *PlasmoDBv10.0*.

	Contiguity	Chimerism	Total	Intergenic	Antisense
PlasmoDBv10.0			5,777		
Cufflinks	81.5%	7.9%	7,065	660	479
Cufflinks RABT	100%	6.7%	9,434	660	474
Genome-guided Trinity	57.1%	3.8%	43,816	8,260	11,070
Genome-guided Trinity RABT	100%	2.5%	21,182	5,839	7,234

*RABT = Reference annotation based transcript assembly

To further benchmark Cufflinks assembly performance in *P. falciparum*, we compared the expression properties, GC content, and length of Cufflinks-assembled transcripts to those of previous annotations. Towards this end, we paired 5727 and 7736 assembled transcripts with *PlasmoDBv10.0* annotated transcripts in the unassisted and assisted Cufflinks assemblies, respectively. We then calculated the correlation between paired expression profiles, finding a median correlation of 0.98 and 0.99 for unassisted and assisted transcripts, respectively. This led us to conclude that analyzing assembled transcript expression profiles was essentially interchangeable with analyzing annotated transcript expression profiles. We did, however, note a shift towards lower FPKM (fragments per kilobase of exon per million fragments mapped) expression level and lower GC content for assembled transcripts. This was largely because assembled transcripts were generally longer and included unannotated, likely un-translated regions (UTRs) with less read support and reduced GC content as compared to coding regions [**Supplementary Table 3.4**, **Supplementary Figure 3.5**]. We selected the annotation-assisted Cufflinks transcriptome for further analyses, unless otherwise noted, as it represented the most complete *P. falciparum* transcriptome.

In sum, annotation-assisted Cufflinks assembly predicted 9434 transcripts, including 660 unannotated intergenic transcripts (647 unique loci) and 474 antisense transcripts (467 unique loci) [**Figure 3.1A**, **Figure 3.3A**, **see GSE57439**]. The 467 unique antisense loci overlapped 462 annotated genes in an approximately 1:1 ratio. This encompassed transcription of at least 73% of the *P. falciparum* genome, a 13% increase compared to annotation alone, and included the prediction of high-confidence antisense transcription from 8% of annotated genes. Annotation-assisted Cufflinks assembly also predicted 2134 novel splice-junctions. On the other hand, Cufflinks assembly without annotation rediscovered 6918 out of 8537 annotated splice-junctions

(81%) and, as noted above, predicted contiguous transcripts spanning 4707 out of 5777 annotated transcripts (81.5%) [**Figure 3.3A**]. These high rediscovery rates suggested that we may have annotated the majority of *P. falciparum* intergenic lncRNAs and antisense RNAs using deeply sequenced, blood stage samples. Of course, further strand-specific sequencing studies of the sexual blood and mosquito stages will be necessary to identify and characterize the full range of non-coding transcripts in *P. falciparum.*

Coding Region Identification

We next explored whether the 1134 previously unannotated transcripts had coding potential, and found that at least 98.5% represented bona fide non-coding RNAs. Using TransDecoder [72], we predicted putative coding regions in 5213 out of the 5229 (99.7%) possible protein-coding transcripts, but in just seven out of the 660 intergenic transcripts (1.1%) and eleven out of the 474 antisense transcripts (2.3%) [**Figure 3.1A**]. These proportions of putative coding regions in our candidate lncRNA sets did not significantly differ from the proportions that TransDecoder predicted in random regions (97 out of 6600 random intergenic regions, Fisher's exact test, p-value = .493; 57 out of 4740 random antisense regions, Fisher's exact test, p-value = .053). Moreover, we found no precedent for overlapping genes in *P. falciparum* [15]. Given this body of data and the small proportion that the ambiguous transcripts represented in their respective data sets, we retained but noted these transcripts for further investigation [**see GSE57439**].

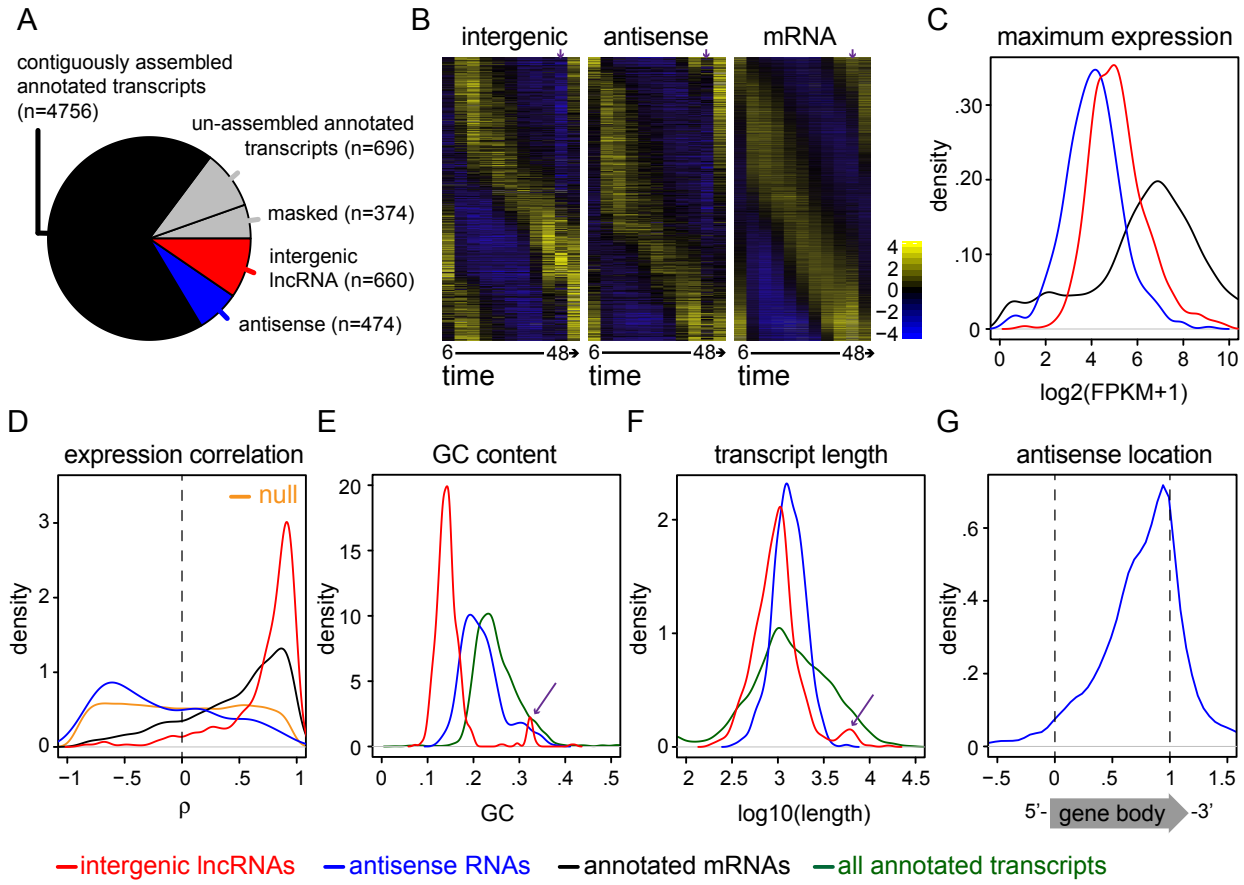


Figure 3.3 | Characterization of 1134 unannotated *P. falciparum* lncRNAs reveals global trends as well as intriguing outliers. (A) Without annotation assistance, at least 4707 out of 5777 (81.5%) annotated transcripts could be contiguously assembled in our blood stage samples. 696 annotated transcripts could not be contiguously assembled, and we excluded 374 short and/or structural RNAs from assembly. Given this high reassembly rate of known transcripts, it is possible that the 660 intergenic lncRNAs and 474 antisense RNAs described here represent the majority of lncRNAs transcribed in *P. falciparum*. **(B)** Comparative inspection of non-clustering heatmaps showed that predicted lncRNAs were developmentally regulated in a similar periodic fashion to annotated mRNAs. However, it was also apparent that a subset of lncRNAs strongly peaked in expression during parasite invasion, and that there was a paucity of antisense expression during parasite invasion. The 48 hpi invasion time-point is indicated with purple arrows. Transcripts are ordered by their angular position in the MDS plot of transcript expression profiles, and samples are ordered by time. Mean-centered expression is in units of log2(FPKM+1). **(C)** The distribution of maximum expression levels for each transcript class suggested that both intergenic lncRNAs (red) and antisense RNAs (blue) were robustly expressed, albeit they typically reached lower maximum expression levels than annotated mRNAs (black).

Figure 3.3 (Continued) | Characterization of 1134 unannotated *P. falciparum* lncRNAs reveals global trends as well as intriguing outliers. (D) Pearson correlation during the 56-hour time course between 50,000 random mRNA gene pairs (orange) as compared to 5251 mRNA-neighboring gene pairs (black), 498 intergenic lncRNA-neighboring gene pairs (red), and 445 antisense-sense gene pairs (blue). To be consistent, we defined the neighboring gene used in both the mRNA and intergenic lncRNA pairings as the more correlated neighboring mRNA. **(E)** The distribution of GC content for each transcript class indicated that intergenic lncRNAs (red) and antisense RNAs (blue) typically had lower GC content than annotated transcripts (green), though a handful of intergenic lncRNAs had unusually high GC content (purple arrow). **(F)** The distribution of transcript length for each transcript class showed that intergenic lncRNAs (red) and antisense RNAs (blue) were comparable in length to annotated transcripts (green), with the average of each class being longer than 1 kb. Markedly long intergenic lncRNAs (> 4 kb) are indicated with a purple arrow. **(G)** Plotting the normalized distribution of antisense RNAs relative to annotated gene bodies revealed a 3' tail-to-tail bias.

Differential Expression and Regulation of the P. falciparum Transcriptome

We leveraged Cuffdiff to compute the differential expression and regulation of our annotation-assisted transcriptome across developmental stages [71]. In brief, we compared samples harvested approximately eight hours apart during the 56-hour time course, and estimated biological variation across the blood stage using biological replicate samples. This approach is unprecedented in *P. falciparum;* previous studies have either not assessed differential expression or have arbitrarily compared samples without biological replication or respect for time course topology [24,43-45]. Employing these specifications and excluding chimeric results, we found 3815 differentially expressed genes, 127 alternative splicing events, and 81 cases of alternative promoter usage in the annotation-assisted transcriptome [**Figure 3.1A**, **see GSE57439**]. This included 354 significant intergenic lncRNA loci (out of 647) and 69 significant antisense loci (out of 467).

LncRNA Transcript Properties

After ensuring data integrity, including validating the non-coding nature of unannotated transcripts, we set out to characterize lncRNA transcript properties. Towards this end, we first compared the expression periodicity of lncRNA transcripts to that of annotated mRNA transcripts, as stage-specific expression is likely to correlate with function. Indeed, when we visualized the expression of each transcript class in a non-clustering heatmap, we found a similar pattern of developmental regulation for both lncRNAs and mRNAs [**Figure 3.3B**], although lncRNAs typically reached lower maximum expression levels than mRNAs [**Table 3.2**, **Figure 3.3C**]. Notably, the lower expression regime of *P. falciparum* lncRNAs was in agreement with previous genome-wide studies of lncRNA characteristics in humans and model organisms [76,77]. Taken together, this global analysis revealed the remarkable similarity between lncRNA and mRNA expression cascades during blood stage development and suggested stage-specific roles for *P. falciparum* lncRNAs.

Table 3.2 | Global properties of *P. falciparum* lncRNAs include reduced expression, length, GC content, and splicing as compared to annotated transcripts. We calculated the average maximum FPKM and average average FPKM across each transcript class during the 56-hour time course. Average length and average GC content reflect exonic sequence only. Annotated transcript properties refer to *PlasmoDBv10.0* transcript models.

	LncRNA	Antisense	Annotated
Average maximum FPKM	56	25	469
Average average FPKM	18	8	164
Average Length	1218	1413	2197
Average GC content	15.0%	21.8%	25.4%
Single exon rate	93.5%	89.5%	47.8%
Maximum exon count	3	5	34

Additionally, this visual approach highlighted two intriguing lncRNA expression profile deviations during RBC rupture and parasite invasion [purple arrows in **Figure 3.3B**]. Upon close inspection of the intergenic lncRNA expression profiles shown in **Figure 3.3B**, we noted that a subset of intergenic lncRNAs strongly peaked in expression during the 48 hpi invasion time-point. We found that this subset included all members of the family of telomere-associated lncRNA-TAREs that we previously identified [42]. Second, upon close inspection of the antisense RNA expression profiles shown in **Figure 3.3B**, we noted a paucity of antisense transcription during parasite invasion. In fact, we calculated that out of the 35% of antisense RNAs (166) increasing in expression between 36-44 hpi, 72% dropped in expression during parasite invasion and then increased in expression afterwards. A similar percentage of annotated mRNA transcripts (27% or 1435) increased in expression between 36-44 hpi, but only 19% exhibited the invasion-specific expression drop (Fisher's exact test, p-value < .0001).

We next investigated the correlation properties of *P. falciparum* lncRNAs and annotated mRNAs, as positive or negative correlation between lncRNAs and neighboring genes may indicate a regulatory relationship [52,78,79]. Specifically, we compared the expression correlation between randomly sampled mRNAs (location-independent null) to that of the following location-dependent gene pairs: (1) annotated mRNAs and their more correlated neighboring mRNA, (2) intergenic lncRNAs and their more correlated neighboring mRNA, and (3) sense-antisense partners. We observed significantly more positively correlated intergenic lncRNA-neighbor pairs and mRNA-neighbor pairs than random mRNA pairs (Wilcoxon rank sum p-value < 2.2e-16 in both cases) [**Figure 3.3D**]. On the other hand, we found that sense-antisense partners exhibited an entirely different expression correlation trend. Namely, we

observed significantly more negatively (or anti-) correlated sense-antisense pairs than random mRNA pairs (Wilcoxon rank sum p-value = 3.834e-11) [**Figure 3.3D**].

Finally, we found that intergenic lncRNA-neighbor pairs were significantly more positively correlated than mRNA-neighbor pairs (Wilcoxon rank sum p-value < 2.2e-16) [**Figure 3.3D**]. Given this feature, we pursued numerous additional analyses of intergenic lncRNA-neighbors and mRNA-neighbors to explore whether positive correlation may be dependent on orientation and/or genomic distance between neighboring loci. In brief, we found that both lncRNAs and mRNAs had a significantly more correlated neighbor (Wilcoxon signed-rank test p-value < 2.2e-16 for both lncRNAs and mRNAs), that the distance between intergenic lncRNA-neighbor pairs was not particularly indicative of higher correlation (ρ = -.25), and that lncRNAs and mRNAs were located at comparable distances from other annotated mRNAs (1576 bp versus 1585 bp, respectively) [**Supplementary Figure 3.6**]. In terms of orientation, we found that expression correlation was equally distributed for tandem (→ →) and divergent (← →) intergenic lncRNA-neighbor pairs, although convergent (→ ←) pairs were markedly less frequent and behaved similarly to mRNA-neighbor pairs (Wilcoxon rank sum p-value = 0.3607) [**Supplementary Figure 3.6**]. Taken together, our results indicated a role for genomic architecture in *P. falciparum* lncRNA and mRNA expression dynamics, as has been recently described by Ay et al. using three-dimensional modeling of the *P. falciparum* genome [36]. However, we did not find genomic distance between neighboring loci to be a strong indicator of expression correlation. Moreover, our results indicated additional complexity, as mRNAs, intergenic lncRNAs, and antisense RNAs had significantly different expression correlation properties with neighboring loci.

We next considered the GC content and length of lncRNA transcripts. The GC content of intergenic lncRNAs was generally lower than that of antisense RNAs, which was lower than that of annotated transcripts [**Table 3.2, Figure 3.3E**]. This was not surprising given the higher GC content of coding sequences, ribosomal RNA, and transfer RNA in the *P. falciparum* genome [15]. In terms of transcript length, both lncRNA classes were quite long, with the average length of intergenic lncRNAs, antisense RNAs, and annotated transcripts being 1218 bp, 1413 bp, and 2197 bp, respectively [**Table 3.2, Figure 3.3F**]. The small subset of relatively GC-rich (> 29%) intergenic lncRNAs generally corresponded to the subset of relatively long intergenic lncRNAs (> 4 kb), and included all members of the telomere-associated lncRNA-TARE family, whose high GC content and length we previously characterized [arrows in **Figure 3.3E** and **Figure 3.3F**] [42]. The only two unannotated transcripts with greater than 40% GC content shared 82% pairwise sequence identity, and they were both situated between *var* pseudogenes and *PHISTB* genes. TransDecoder predicted a coding region in one of these transcripts, and given their high GC content and sequence similarity, we reasoned that both of these transcripts likely represented unannotated pseudogenes.

Finally, we considered the relative location of antisense RNAs within annotated gene bodies and the splicing properties of lncRNAs. This revealed that *P. falciparum* antisense RNAs largely overlapped tail-to-tail with annotated genes, a property that has been described in previous viral, prokaryotic, and lower eukaryotic genome-wide studies [**Figure 3.3G**] [80]. Specifically, the vast majority of *P. falciparum* antisense RNAs initiated transcription downstream of annotated gene bodies and tended to terminate transcription towards the 3' end of gene bodies [**Supplementary Figure 3.7**]. In terms of splicing, we found that lncRNA transcripts were more likely to be single exon than annotated transcripts, although we also

identified several multi-exonic lncRNAs. We calculated that 93.5% and 89.5% of predicted intergenic lncRNAs and antisense RNAs were single exon, respectively, versus 47.8% of annotated transcripts [**Table 3.2**].

Notable LncRNAs

Based on the diverse characteristics examined above, we searched for *P. falciparum* lncRNAs with intriguing properties. For example, we found that a putative Apicoplast RNA methyltransferase precursor (*Pf3D7_0218300*) and an Early Transcribed Membrane Protein (*ETRAMP; Pf3D7_0936100*) transcribe multi-exonic antisense RNAs across their full gene bodies [**Figure 3.4A** and **Figure 3.4B**]. Expression of the Apicoplast RNA methyltransferase precursor sense-antisense pair was not particularly correlated ($\rho = .20$), while expression of the *ETRAMP* sense-antisense pair was moderately anti-correlated ($\rho = -.50$) [**Figure 3.4C** and **Figure 3.4D**]. Interestingly, *ETRAMP* antisense transcription was substantially higher than *ETRAMP* sense transcription, reaching a maximum FPKM of 550 in early stages. This was the highest expression level observed for predicted *P. falciparum* antisense RNAs at any stage. Both the Apicoplast RNA methyltransferase precursor and *ETRAMP* antisense RNAs also demonstrated the 48 hpi invasion expression drop phenomenon. Namely, both antisense transcript levels increased between 36-44 hpi, dropped during parasite invasion, and then increased. Their sense partners did not exhibit this pattern.

Figure 3.4 | Notable lncRNAs include multi-exonic and telomere-associated transcripts. (A)/(B) Multi-exonic antisense transcripts span an apicoplast RNA methyltransferase precursor (*PF3D7_0218300*) and an *ETRAMP* (*PF3D7_0936100)* gene, respectively. Annotated gene models are shown in dark green and dark blue, and assembled transcript models are shown in light green and light blue. Reads from each 56-hour time course sample mapping to the (-) strand are shown below each horizontal axis in light green, while reads mapping to the (+) strand are shown above each horizontal axis in light blue. Intron reads are shown in purple. Uniqueness of 100mers is plotted in red as a mappability track. **(C)/(D)** Pearson correlation between the *Pf3D7_0218300* sense-antisense pair and *ETRAMP* sense-antisense pair during the 56-hour time course was 0.20 and -0.50, respectively. Notably, *Pf3D7_0218300* and *ETRAMP* antisense transcript levels dropped during parasite invasion, while sense transcript levels did not. Expression is plotted in units of log2(FPKM+1). **(E)** Multi-exonic lncRNAs are encoded in the *PfGDV1* region on chromosome nine, antisense to *PfGDV1* and between *PfGDV1* and *GEXP22*. Refer to **(A)/(B)** for a description of tracks. **(F)/(G)** Pearson correlation between the *PfGDV1* sense-antisense pair was 0.96, while Pearson correlation between the divergent intergenic lncRNA and *GEXP22* pair was 0.46 during the 56-hour time course. Expression is plotted in units of log2(FPKM+1). **(H)** As we have previously described, the telomere-associated repetitive element (TARE) 2-3 region transcribes a family of lncRNA-TAREs, with transcription always proceeding towards the telomere [42]. For example, lncRNA-TARE-2L is transcribed on the left arm of chromosome two. *Pf3D7_0200100* is a subtelomeric upsB-type Pfemp1-encoding *var* gene. Boundaries of the telomere, TAREs 1-5, and Rep20 are shown in purple. See **(A)/(B)** for a further description of tracks. **(I)** Plotting the expression level of 22 lncRNA-TARE family members showed that lncRNA-TARE expression was co-regulated, with maximal firing coinciding with parasite invasion. Expression is plotted in units of log2(FPKM+1). **(J)** Pearson correlation between lncRNA-TARE-2L and the neighboring Pfemp1-encoding *var* gene was -0.09 during the 56-hour time course. Expression is plotted in units of log2(FPKM+1).

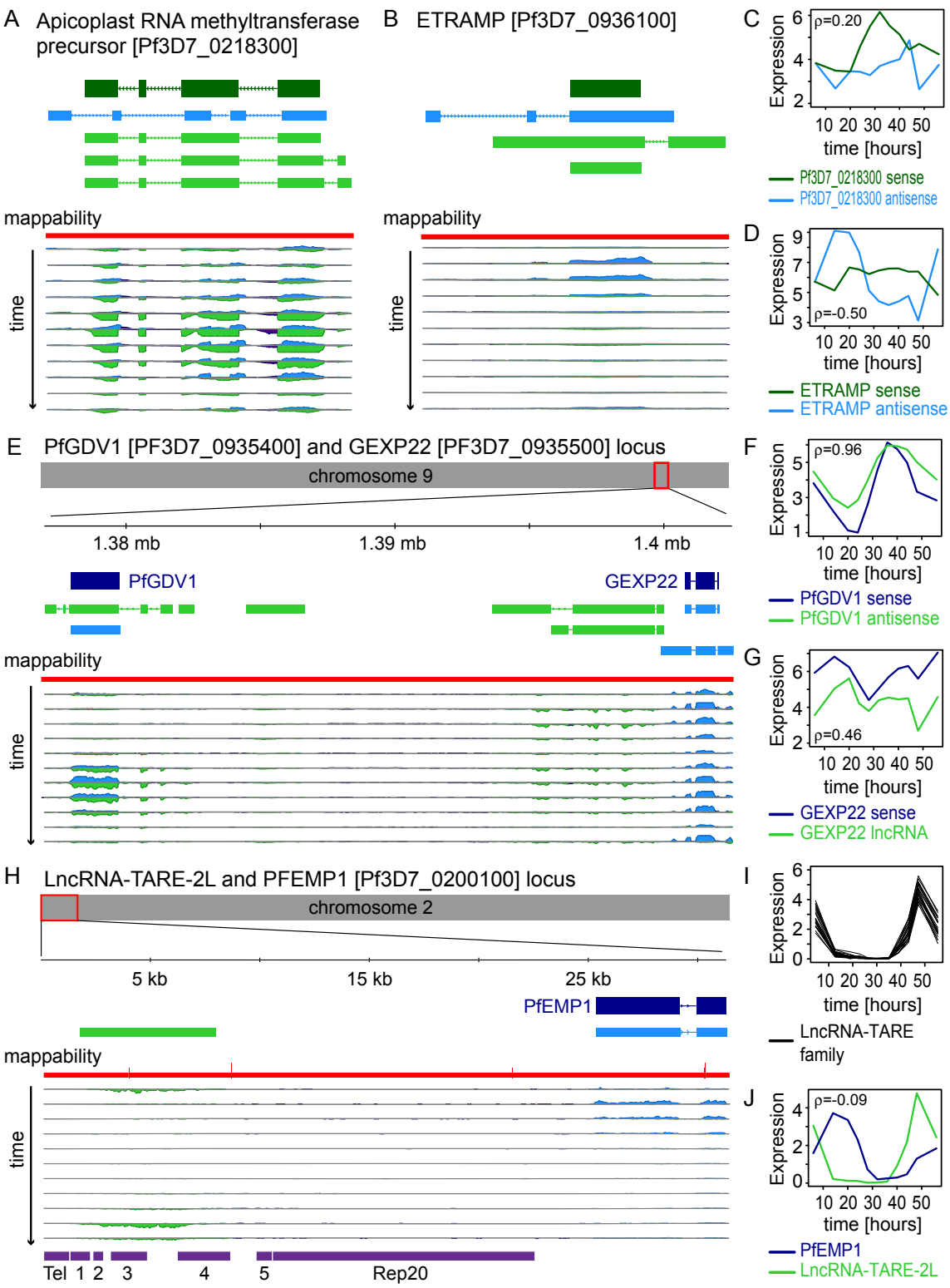


Figure 3.4 (Continued) | Notable lncRNAs include multi-exonic and telomere-associated transcripts.

Remarkably, our approach also revealed that a region on chromosome nine required for early sexual development [33] harbors a highly expressed, developmentally regulated, multi-exonic antisense transcript to *P. falciparum* Gametocyte Development Protein 1 (*PfGDV1; Pf3D7_0935400*), as well as two intergenic lncRNAs downstream of *PfGDV1* [**Figure 3.4E**]. Correlation during the 56-hour blood stage time course between *PfGDV1* sense and antisense transcript levels was the highest of any predicted *P. falciparum* sense-antisense pair ($\rho = 0.96$), with *PfGDV1* antisense transcript levels typically exceeding *PfGDV1* sense transcript levels [**Figure 3.4F**]. This was in sharp contrast to the majority of *P. falciparum* sense-antisense pairs, which displayed a trend towards anti-correlated expression. Notably, while the expression correlation was again high between *PfGDV1* sense-antisense transcript levels in the four biological replicate samples ($\rho = 0.85$), the difference between transcript levels was greater, with the *PfGDV1* antisense transcript reaching a maximum FPKM of 255 [**Supplementary Figure 3.8**]. The nearby, multi-exonic, intergenic lncRNA exhibited only moderately correlated expression to GEXP22 (*Pf3D7_0935500*) and evidence of alternative splicing ($\rho = 0.46$) [**Figure 3.4G**]. In summary, the *PfGDV1* antisense transcript's expression properties, multi-exonic structure, and position relative to other genes made it a clear outlier in the genome.

While we have previously detected and characterized telomere-associated lncRNA-TAREs, the properties of this yet to be annotated lncRNA family again stood out in our analyses [42]. Our results confirmed that lncRNA-TAREs were long, high-GC, and transcribed towards the telomere [**Figure 3.4H**, and arrows in **Figure 3.3E** and **Figure 3.3F**]. We also confirmed that lncRNA-TARE transcription was generally restricted to the expected TARE 2-3 region, although we did find that in one case the entire TARE 1-6 region was transcribed [**Supplementary Figure 3.9**]. To build on our previous results, long, paired-end, uniquely mapped sequencing reads

showed that lncRNA-TARE transcripts likely originated from 22 chromosome ends in our parasite populations. Moreover, the increased time resolution and scope of our samples showed that lncRNA-TARE transcript levels coordinately peaked during parasite invasion [**Figure 3.4I**]. Interestingly, we noted that the expression of previously described, sterile *var* transcripts peaked during parasite invasion as well, but that not all *var* genes produced these non-coding transcripts [**Supplementary Figure 3.3**] [81]. For example, the subtelomeric *var* gene (*Pf3D7_0200100*) neighboring lncRNA-TARE-2L was lowly expressed during the ring stage and did not produce appreciable sterile transcripts [**Figure 3.4H** and **Figure 3.4I**]. Collectively, these findings suggested co-regulated firing of lncRNA-TARE and sterile *var* transcripts during parasite invasion, which could indicate coordinated function.

Discovery of CircRNAs in *P. falciparum*

Given recent reports of circRNAs across eukaryotes and their hypothesized regulatory functions, we searched for evidence of *P. falciparum* circRNAs in our data set as well [46,82,83]. CircRNAs can be distinguished in non-polyA-selected deep sequencing libraries by spliced reads in which sequence ending in a splice donor maps downstream of sequence beginning with a splice acceptor. Following the analysis pipeline and criterion published by Memczak et al., we found 1381 predicted *P. falciparum* circRNAs, between 0.1 and 10 kb long, with at least two unique reads spanning their splice junction [**Figure 3.1A**] [83]. Of these, 273 had five or more unique reads of support (the gold standard being 2 reads) [**Supplementary Table 3.5]**. As compared to the transcriptome-wide results reported in **Table 3.2**, we found that *P. falciparum* transcripts with predicted circRNAs were more highly expressed on average (2646 average maximum FPKM; 791 average average FPKM). Indeed, the circRNA-producing gene

set was enriched for ribosome-related components; ribosomal proteins are typically highly expressed [**Supplementary Table 3.6**].

In contrast to human circRNAs, *P. falciparum* circRNAs were generally predicted to be short, with the majority being less than 200 bp [84]. Only 510 out of 1381 predicted circRNAs with at least two unique supporting reads were predicted to be 200 bp or longer. In the more stringent set of 273 circRNA candidates with at least five unique supporting reads, only 72 were predicted to be 200 bp or longer. We defined circRNA size as the genomic distance between predicted donor site and acceptor site, inclusive of the donor and acceptor site. Thus, this should be read as a maximum size, as circRNAs can span introns, which may be spliced out of the circRNA sequence. In summary, short circRNAs appeared to outnumber longer circRNAs in *P. falciparum* and deserve further attention.

We predicted an intriguing top *P. falciparum* circRNA candidate within the apoptosis-related protein (*ARP;Pf3D7_0909300)*, termed ARP_circRNA [**Figure 3.5A**]. 56 unique reads spanned the predicted splice junction between ARP's exon-4 donor site (GT) and upstream exon-3 acceptor site (AG) [**Figure 3.5B**]. To validate that this non-canonical junction was not the result of a library preparation or sequencing artifact, we reverse-transcribed total RNA and amplified the predicted ARP_circRNA junction from the resulting complementary DNA (cDNA) using PCR and divergent primer pairs. We designed divergent ARP_circRNA primer pairs, as is depicted in **Figure 3.5B**, such that primer pairs could not amplify genomic DNA (gDNA) or cDNA in the absence of the predicted ARP_circRNA splice junction.

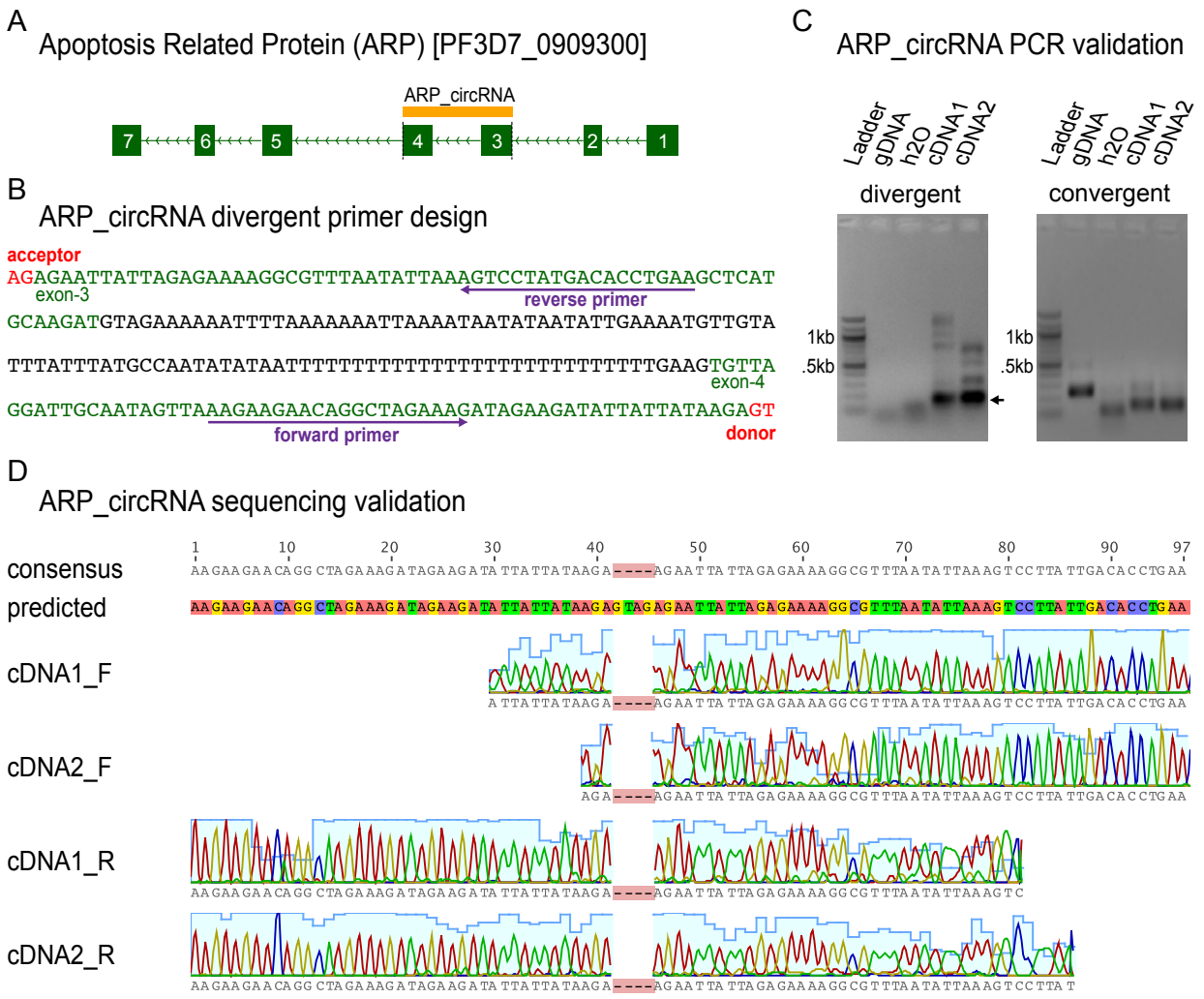


Figure 3.5 | Divergent primers and Sanger sequencing validate circRNA splicing in *P. falciparum*. **(A)** The apoptosis-related protein (*ARP*) encodes a predicted circRNA, termed ARP_circRNA, consisting of *ARP* exon-3 and exon-4 sequence. **(B)** To validate the non-canonical exon-4 donor (GT)/exon-3 acceptor (AG) splice junction in ARP_circRNA, we designed a divergent PCR primer pair. The primer pair is considered to be divergent, rather than convergent, because the reverse primer binds upstream of the forward primer. **(C)** PCR using divergent primers amplified a product of the expected size (161 bp, indicated with an arrow) when the template was cDNA from either time course, but not water or gDNA. The larger products in the divergent cDNA reactions may represent non-specific or rolling-circle reverse transcription products [46]. On the other hand, PCR using convergent primers amplified products of the expected size when the template was cDNA from either time course or gDNA. We confirmed that the smaller product size in the case of convergent cDNA reactions corresponded to intron removal. **(D)** Sanger sequencing of divergent amplicons of the expected size confirmed the ARP_circRNA junction in both time courses. The extra GTAG in the predicted sequence marks the non-canonical ARP_circRNA splice junction (highlighted in red in the consensus sequence).

Our results confirmed that ARP_circRNA represented a bona fide *P. falciparum* circRNA. The ARP_circRNA divergent primer pair produced amplicons of the expected size when the template was cDNA prepared from either biological replicate time course, and did not produce specific amplicons with gDNA or water as the template. On the other hand, the ARP_circRNA convergent primer pair amplified both cDNA and gDNA, with the smaller product size in the cDNA reactions corresponding to intron removal [**Figure 3.5C**]. We further confirmed the identity of the ARP_circRNA divergent and convergent amplicons by Sanger sequencing. Sequence confirmation for the ARP_circRNA divergent amplicon is shown in **Figure 3.5D**, where the GTAG splice donor-acceptor tag is included in the predicted sequence as a marker for the circRNA splice junction.

We used a parallel strategy to validate five out of eight additional *P. falciparum* circRNAs, for a 67% validation rate of high quality candidates [**Supplementary Figure 3.10** and **Supplementary Table 3.7**]. The nine tested candidates had the most read support out of predicted *P. falciparum* circRNAs with either a donor or acceptor site in common with an annotated transcript, were predicted to be larger than 200 bp (genomic distance), and were not predicted within ribosomal genes. Interestingly, while two of the additional validated *P. falciparum* circRNAs were associated with genes of unknown function, two were predicted within rhoptry-related genes, and the final validated candidate was within metacaspase-like protein (MCA2), another gene involved in apoptosis.

3.3 Discussion

In this work we make immense progress towards the functional characterization of *P. falciparum* lncRNAs. We assemble 1134 unannotated *P. falciparum* intergenic lncRNA and antisense transcript structures, predict 1381 *P. falciparum* circRNA candidates (6/9 experimental confirmation rate), compile a comprehensive catalog of transcript properties, summarize global trends, and prioritize a small set of *P. falciparum* lncRNA candidates for immediate investigation. This required development of a state-of-the-art, strand-specific RNA sequencing protocol sensitive to the extreme AT content of *P. falciparum,* unprecedented time-point resolution of sequencing samples, and biological replication. Despite a handful of important studies over the past few years in this space [42-45], including our own pilot study using high-density tiling microarrays [42], the next-generation sequencing advancements, sample set, and computational pipeline that we employ here technologically surpass previous studies.

Our findings have several implications for parasite biology and shed new light on previous *P. falciparum* sequencing studies, especially in regards to the function and expression properties of antisense transcripts. Lopez-Barragan and colleagues suggested that *P. falciparum* genes transcribed bidirectionally (sense and antisense) were more likely to be spliced, and hypothesized that *P. falciparum* antisense RNAs may be involved in the splicing of sense transcripts [43]. However, we found that 43.1% of genes transcribed bidirectionally were single exon, which was not significantly different from the genome-wide rate of 47.8% (Fisher's exact test, p-value = .053). As another example, Siegel et al. predicted that 22% of annotated genes produce antisense RNAs (1247), and did not find global correlation or anti-correlation between sense-antisense pair expression [45]. We identified high-confidence antisense transcription from

8% of annotated *P. falciparum* genes (462), and detected a significant global shift towards anti-correlated sense-antisense pair expression [**Supplementary Figure 3.11**].

The enrichment of negatively correlated, tail-to-tail overlapping sense-antisense transcript pairs aligns with a role for many *P. falciparum* antisense RNAs in transcriptional and/or post-transcriptional regulation of their sense mRNA partners [49]. For example, *P. falciparum* antisense RNAs could function through transcriptional interference, as has been extensively studied in *Saccharomyces cerevisiae* [80,85-87]. In the transcriptional interference model, antisense transcription interferes with sense transcription through either polymerase collisions or alternative mechanisms. As an alternative or additional model, antisense-mediated transcriptional suppression is also possible and has been described in human studies [88-90]. In antisense-mediated transcriptional suppression, antisense RNAs act as epigenetic silencers, catalyzing local heterochromatin formation. *P. falciparum* antisense RNAs may also act post-transcriptionally on sense transcripts by inhibiting their nuclear export or translation [49]. In support of this notion, peptide nucleic acids (PNAs) designed as specific antisense molecules have been recently shown to regulate both transgenic and endogenous *P. falciparum* genes in a sequence-specific, post-transcriptional manner [91]. Albeit controversial, numerous studies from the 1990s also reported on antisense oligonucleotide-mediated silencing of target genes in *P. falciparum* [92,93]. Taken together, the often antagonistic interplay between sense and antisense transcript levels in *P. falciparum* aligns with known regulatory mechanisms of antisense RNAs in humans and model organisms, and is likely to be functionally significant.

Searching our catalog for *P. falciparum* lncRNAs with unique properties revealed that an essential protein in early gametocyte development, *PfGDV1,* has a highly and coordinately expressed, multi-exonic antisense counterpart*,* as well as multiple neighboring intergenic

lncRNAs. Though the regulation and mechanism of early gametocyte development events remain largely unknown, Eksi et al. have shown that *PfGDV1* complementation restores gametocytogenesis in *PfGDV1*-null parasites, and that episomal *PfGDV1* over-expression increases gametocytemia in wild-type parasites [33]. This suggests that endogenous *PfGDV1* expression levels might correlate with gametocytemia as well, and that silencing the endogenous *PfGDV1* locus could disrupt transmission. Notably, Kafsack et al. have also shown that a member of the ApiAP2 transcription factor family is involved in early gametocyte development [37]. However, loss of this factor did not affect *PfGDV1* transcript levels [37], suggesting that the *PfGDV1* locus may be subject to additional regulatory complexity. In light of these recent findings, and given that future strategies to block malaria transmission largely hinge on understanding the development of the transmissable *P. falciparum* sexual stages [12-14], the lncRNA transcripts in the *PfGDV1* region warrant immediate functional follow-up. Specifically, we hypothesize that the antisense RNA to *PfGDV1* may play a major regulatory role in early gametocyte commitment events, perhaps similar to *S. cerevisiae* antisense-mediated entry into meiosis [52,53].

We previously described and hypothesized that a family of telomere-associated lncRNA-TARE transcripts is involved in telomere maintenance and/or subtelomeric *var* gene regulation [42,94]. Supporting such nuclear roles, Siegel et al. have recently reported significant enrichment of telomere-associated *P. falciparum* lncRNAs in the nuclear, as opposed to total and cytoplasmic, *P. falciparum* RNA fractions [45]. Adding to the lncRNA-TARE family profile, we have now shown that lncRNA-TARE transcripts are maximally expressed during parasite invasion, along with the sterile, non-coding *var* transcripts transcribed from bidirectional *var* intron promoters [81]. Interestingly, when parasites invade RBCs pre-loaded with episomal

Pfemp1-encoding *var* genes, these episomal *var* genes are not silenced during the first ring stage. *Var* gene silencing is, however, somehow established by the subsequent ring stage and then remains the default status [95]. While it has been hypothesized that passage through S-phase is the critical requirement for *var* gene silencing, we suggest that expression of non-coding transcripts during parasite invasion may play an important regulatory role as well [96,97]. High-resolution studies of chromatin mark dynamics over episomal and endogenous *var* loci during parasite invasion may help to resolve these models.

Finally, this work represents the discovery and validation of at least seven bona fide *P. falciparum* circRNAs, as well as the prediction of hundreds more. With the majority of predicted *P. falciparum* circRNAs being shorter than the average library fragment size, and, as such, likely selected against during library preparation, a focused sequencing effort using circRNA-specific protocol adjustments may reveal even more abundance of this intriguing class of non-coding RNAs. In support of this notion, Wang et al. recently mined the Siegel et al. RNA sequencing data set and validated two low-abundance *P. falciparum* circRNAs [45,46]. As we did not find evidence for either of these candidates in our study, this further hints that library preparation and/or bioinformatic processing may be tunable to capture additional *P. falciparum* circRNAs. Alternatively, as our sample sets differed, these circRNAs may be highly linked to the phenotypic state of the parasite.

Notably, Memczak et al. and others have hypothesized that circRNAs may act as microRNA sponges [83,84]. In this model, the circRNAs behave as competitive inhibitors, sponging up microRNAs to reduce the microRNA pool available to bind messenger RNA targets. However, it is highly unlikely that microRNA binding is involved in *P. falciparum* circRNA function, as *P. falciparum* does not encode microRNAs nor microRNA processing

machinery, and the only reported, non-canonical function for human microRNAs in *P. falciparum* does not involve microRNA binding either [41,98]. Taken together, our discovery of hundreds of circRNAs in an organism lacking RNA interference points to potential additional roles for circRNAs across eukaryotes.

3.4 Materials and Methods

Parasite Culture and Sample Harvesting

We harvested a total of fifteen samples from two independent *P. falciparum* blood-stage time courses. In each experiment, we cultured a freshly thawed *P. falciparum* strain 3D7 clone (ATTC) in human RBCs from healthy anonymous donors (Research Blood Components, Boston, MA) using standard methods [11]. We maintained cultures at 4% hematocrit and supplemented RPMI-HEPES medium with 5% human serum (O+) and 5% Albumax II (Gibco). We synchronized cultures using multiple 5% sorbitol solution treatments [99], and expanded cultures to accommodate harvesting of at least 50 mL of culture at each planned time-point (time course 1: every ~4 hours for 56 hours; time course 2: ~4 hours before and after the ring to trophozoite and trophozoite to schizont morphological stage transitions). We centrifuged 50 mL aliquots of harvested culture at 2400 rpm in a Sorval RT6000B to obtain at least 2 mL of packed RBCs per time-point. For time course 1, we stored packed RBCs in 15 mL of Buffer RLT (with BME added) at -80°C prior to RNA extraction. For time course 2, we lysed packed RBCs using a .05% saponin solution, washed liberated parasites using phosphate-buffered saline (pH 7.4), pelleted parasites at 13.2 RPM in a micro-centrifuge, and stored parasites in 1 mL of TRIZOL reagent at -80°C prior to RNA extraction.

Total RNA Extraction

For time course 1, we thawed RBC samples stored in Buffer RLT, added one volume of 70% ethanol, and immediately loaded the mixture onto RNeasy Midi columns (Qiagen). For time course 2, we thawed parasite samples stored in TRIZOL reagent, performed TRIZOL-chloroform extraction, and immediately applied the aqueous layer to RNeasy Mini columns

(Qiagen). During both RNeasy Midi and Mini RNA extraction procedures, we performed the optional on-column DNase I digestion for thirty minutes to remove genomic DNA. We stored isolated total RNA aliquots at -80°C with 1 unit/uL RNaseOUT (Invitrogen), and validated RNA quality using an Agilent Bioanalyzer RNA 6000 Pico Kit.

Strand-specific, Non-polyA-selected, Library Preparation and Sequencing

We began library preparation with a second DNase treatment (Ambion TURBO DNase) using 20 units of SUPERase-In (Ambion) and 40 units of RNaseOUT (Invitrogen) to protect RNA from degradation. Each DNase reaction was incubated at 25°C for 30 minutes followed by 1.8X RNAclean SPRI bead purification (Agencourt). Second, we used a Human/Mouse/Rat Ribo-Zero Magnetic Kit (Epicentre) to deplete 18S and 28S rRNA from DNase-treated total RNA. We used 3.5-5 ug of DNase-treated total RNA for all samples except T6, T14, and TT8. In these cases, we used .4 ug, 1 ug, and .7 ug of DNase-treated total RNA, respectively. Furthermore, for T6 we mixed 4 and 8 hpi total RNA 1:1, and for T14 we mixed 12 and 16 hpi total RNA 1:1. Third, we fragmented rRNA-depleted RNA at 85°C for 8 minutes using Mg^{2+} Fragmentation Buffer (New England Biolabs), followed by 1.8X RNAclean SPRI bead purification (Agencourt). Fourth, we reverse-transcribed the fragmented RNA using SuperScript III (Invitrogen), 200 ng/uL of freshly prepared Actinomycin D (Sigma-Aldrich), 3 ug of 76% AT-biased random hexamers (Integrated DNA Technologies), and a gradually ramping-up thermocycler program. Specifically, we set the ramp speed of a PTC-225 DNA Engine Tetrad (MJ Research) to .1°C/second and used the following program: 5°C, 10°C, 15°C, 20°C, 25°C, 30°C, 35°C for 5 minutes each, 42°C for 30 minutes, 45°C, 50°C, 55°C for 10 minutes each. We cleaned up first strand synthesis (FSS) reactions using both Micro Bio-Spin P-30 RNase-free columns (Bio-Rad) and 1.8X RNAclean SPRI bead purification (Agencourt). Fourth, we

performed second strand synthesis (SSS) using a biased dACG-TP/dU-TP mix (Fermentas), 10 units of *E. coli* DNA ligase (Invitrogen), 160 units of *E. coli* DNA polymerase (Invitrogen), and 2 units of *E. coli* RNase H (Invitrogen), followed by 1.8X AmpureXP SPRI bead purification (Agencourt). Fifth, we used an Illumina series KAPA Library Preparation Kit (Kapa Biosystems) and barcoded Y-adapters developed by the Broad Institute to end repair, A-tail, and ligate adapters to each library. We added adapters in approximately 15-fold excess of library targets, and removed un-ligated adapters and adapter-dimers using 1.0X AmpureXP SPRI bead purification (Agencourt). Sixth, we digested the dUTP-marked second strand at 37°C for 30 minutes, followed by 25°C for 15 minutes, using Uracil-Specific Excision Reagent (USER) enzyme (New England Biolabs). Seventh, we amplified libraries for as few cycles as necessary using a KAPA Real-Time PCR Library Amplification Kit (Kapa Biosystems) and PCR primers developed by the Broad Institute. Each library except for T6, T14, and TT8 required only four PCR cycles, while T14 and TT8 required eight PCR cycles, and T6 required twelve PCR cycles. Following a 2 minute denaturation step at 98°C, we cycled libraries using an ABI 7900 Real-Time PCR machine and the following 2-step program: (1) denaturation at 98°C for 20 seconds, (2) annealing and extension at 55°C for 190 seconds.

Finally, we quantified libraries using a KAPA Library Quantification Kit (Kapa Biosystems) and an Agilent Bioanalyzer High Sensitivity DNA Kit. We combined barcoded libraries into two pools, and sequenced each pool on an Illumina Hiseq 2000 machine (one lane per pool) using 101-bp, paired-end read technology. We prepared the fifteen libraries used in this study in parallel (except for real-time amplification). We used the FASTX-Toolkit v0.0.13 to assess raw read quality. See **Supplementary Figure 3.1** and **Supplementary Protocol 3.1** for a library preparation flowchart and a detailed library preparation protocol, respectively. The data

discussed in this publication have been deposited in NCBI's Gene Expression Omnibus [100] and are available through GEO Series accession number GSE57439 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57439).

Read Processing, TopHat Read Alignment, and Aligned Read Benchmarking

We trimmed the last base from reads passing Illumina filtering (PF) using Picard release 1.94 (SamToFastq). We then aligned reads from each sample to the *PlasmoDBv10.0* reference genome using TopHat v2.0.9 and the following parameters: -r 300 –mate-std-dev 100 –library-type fr-firststrand –i 70 –I 5000 –read-mismatches 0 –segment-mismatches 0 –max-segment-intron 5000 –max-coverage-intron 5000 –b2-very-sensitive –read-gap-length 0 –read-edit-dist 0 –read-realign-edit-dist 0 –max-deletion-length 0 –max-insertion-length 0 –max-multihits 2 –no-mixed –no-discordant [63]. These parameters specified an average fragment size of 300 bp with a standard deviation of 100 bp, strand-specific reads prepared using the dUTP method, and an expected intron size of 70-5,000 bp. Also, to allow only perfect read pair mapping in the expected orientation, and to report only two alignments for non-unique read pairs.

To remove non-unique reads, we then used SAMtools v0.1.19 (view –q 10) to remove read pairs with a mapping quality of 10 or less, meaning at least a 10% chance that the read pair truly came from somewhere else in the genome. We used SAMtools v0.1.19 (flagstat) to compile the unique read alignment rate per sample. Subsequently, we used RNA-SeQC hosted by Genepattern to calculate the strand-specificity, exonic rate, intronic rate, intergenic rate, gap rate, and coefficient of variation of the top 1000, 2000, and 4000 expressed genes for each set of aligned reads [64,101]. We also used RSeQC to calculate the rRNA rate and insert size of aligned reads, and the GC content of rRNA-filtered, aligned reads [65]. We used the FASTX-Toolkit v0.0.13 to assess aligned read quality.

We generated the mappability track displayed (in red) in the Array Studio version 7.0 Genome Browser (OmicSoft Corporation) screen shots using the following procedure: (1) We shredded the *PlasmoDBv10.0* genome (excluding mitochondrial and apicoplast contributions) into every possible 100mer sequence. (2) We used the R-project Biostrings package to search for each 100mer sequence in the *PlasmoDBv10.0* genome. (3) We scored the mappability of each 100mer by tallying the number of exact matches possible. (4) Lastly, we viewed the minimum mappability score of 100mers along the genome, such that deviations above the baseline (score of one) indicated regions where 100-bp *single-end* reads could not uniquely align. We calculated that 96.53% of all possible 100mers in the *PlasmoDBv10.0* genome are unique. We used MUSCLE to compute the pairwise sequence identity between *CLAG3.1* and *CLAG3.2* coding regions (96.4%) [102].

Cufflinks Assembly

We used Cufflinks v2.1.1 to assemble aligned reads from each sample into transfrags using the following parameters: –library-type fr-firststrand –max-intron-length 5000 –overlap-radius 1 –min-isoform-fraction .25 –pre-mrna-fraction .25 –min-frags-per-transfrag 50 –trim-3-dropoff-frac .2 –frag-bias-correct [73]. These parameters specified strand-specific reads prepared using the dUTP method and a maximum intron size of 5,000 bp. Also, to not merge transfrags, to filter minor isoforms less than 25% as abundant as the major isoform, to ignore intronic alignments if not as abundant as specified, to require at least fifty aligned reads per assembled transfrag, to trim the 3' end of assembled transfrags at 20% of average coverage, and to run the built-in bias correction algorithm prior to estimating transcript abundance. We also specified a set of 374 short (<300 bp) and/or structural RNAs (all *PlasmoDBv9.3* transcripts with class code: tRNA, rRNA, or snoRNA) to exclude from assembly (and abundance estimation), such that

analyses relying on expression profiles correspond to mRNAs. We used Cuffmerge to parsimoniously merge assembled transfrags from each sample into a final transcriptome assembly, with or without incorporating *PlasmoDBv10.0* annotated transcript models (RABT) [71].

Genome-guided Trinity Assembly

We merged all trimmed reads passing Illumina filtering (PF), aligned merged reads to the *PlasmoDBv10.0* genome using GSNAP release 2013-11-27, and then used scripts from Trinity release 2013-11-10 to assemble a genome-guided *P.falciparum* transcriptome (PASA release 2013-09-07, GMAP release 2013-11-27, Blat v34) [72]. We specified a maximum intron size of 5,000 bp and strand-specific reads prepared using the dUTP method where appropriate. We also parsimoniously merged the genome-guided Trinity assembly with *PlasmoDBv10.0* annotated transcript models (using Cuffmerge [71]) to yield a Genome-guided Trinity RABT assembly.

Assembly Performance Assessment

To assess Cufflinks and genome-guided Trinity assembly performance, we used Cuffcompare to calculate the total number of assembled transcripts, the number of intergenic transcripts, and the number of antisense transcripts [71]. We also transformed assembly GTF files into BED12 files, and used BEDTools v2.18.2 to investigate assembly contiguity and chimerism [103]. We defined contiguity as the rate of *PlasmoDBv10.0* annotated transcripts covered by one assembled transcript over at least 90% of the annotated transcript exonic length in the correct orientation. We defined chimerism as the rate of assembled transcripts that span more than one *PlasmoDBv10.0* annotated transcript in the correct orientation. For contiguity rate, we used the following BEDTools command: < bedtools intersect –split –s –f .90 –wa –wb –a *PlasmoDBv10.0*.bed12 –b Assembly.bed12 >. For chimerism we used the following BEDTools

command: < bedtools intersect –split –c –s –a Assembly.bed12 –b *PlasmoDBv10.0*.bed12 >, followed by brief post-processing.

We next paired 5727 and 7736 assembled transcripts from the Cufflinks and Cufflinks RABT transcriptomes, respectively, with *PlasmoDBv10.0* annotated transcripts, and computed the Pearson correlation between paired expression profiles. Previously unannotated assembled transcripts (including antisense) were naturally excluded from pairings. We calculated a median correlation of .98 and .99 for Cufflinks and Cufflinks RABT paired expression profiles, respectively. We also compared the following properties of paired Cufflinks transcripts, paired Cufflinks RABT transcripts, and *PlasmoDBv10.0* annotated transcripts: average maximum FPKM, average average FPKM, average length, average GC content, single exon rate, and maximum exon count. We only considered exonic sequence in transcript length and GC content calculations [**Supplementary Table 3.4** and **Supplementary Figure 3.5**].

Coding Region Prediction

We used TransDecoder release 2013-11-17 to predict coding regions in the Cufflinks RABT transcriptome [72]. TransDecoder identifies regions in spliced transcript models that likely encode peptides greater than 100 amino acids based on (1) a match to a Pfam domain above the noise threshold or (2) a log-likelihood score of a Markov model for coding DNA that is greater than zero and greatest when calculated in the predicted open reading frame. Using this approach, we predicted coding regions in 5213 out of 5229 assembled protein-coding transcripts at least 100 amino acids long (99.7%), versus just 7 out of 660 intergenic transcripts (1.1%) and 11 out of 474 antisense transcripts (2.3%).

To estimate the background rate of coding potential detected by TransDecoder, we used BEDTools v2.18.2 (shuffle) to randomly generate 6600 size-matched intergenic intervals and 4740 size-matched antisense intervals [72,103]. Following minor post-processing, we paired the shuffled regions with *PlasmoDBv10.0* transcript models (to train TransDecoder) and assessed coding potential prediction. TransDecoder predicted coding regions in 97 out of the 6600 random intergenic regions (1.5%) and 57 out of the 4740 random antisense regions (1.2%). We concluded from this analysis that the rate of coding potential in the lncRNA transcript sets was not significantly different from the background rate in size-matched shuffled regions (Fisher's exact test, p-value = .493 for intergenic regions; Fisher's exact test, p-value = .053 for antisense regions).

Differential Expression and Regulation

To assess differential expression and regulation of both the *PlasmoDBv10.0* and the Cufflinks RABT transcriptomes, we used Cuffdiff v2.1.1 and the following parameters: –library-type fr-firststrand –time-series –min-reps-for-js-test 1. These parameters specified strand-specific reads prepared using the dUTP method, time course samples, and to not require replication of all conditions. We again specified a set of 374 short (<300 bp) and/or structural RNAs (all *PlasmoDBv9.3* transcripts with class code: tRNA, rRNA, or snoRNA) to exclude from abundance estimation, such that analyses relying on expression profiles correspond to mRNAs.

We instructed Cuffdiff to compare samples from the 56-hour time course harvested approximately 8 hours apart (T6 vs. T14, T14 vs. T24, T20 vs. T28, T24 vs. T32, T28 vs. T36, T32 vs. T40, T36 vs. T44, T40 vs. T48, T48 vs. TT8) and to estimate biological variation across the blood stage using the four replicated conditions [71]. Specifically, we considered the correlation matrix between sample profiles and paired R with T14 and T20, ET with T28, LT

with T32, and S with T40. We then supplied these replicated conditions to Cuffdiff as samples in addition to specifying the desired sample comparisons described above. Finally, we used the Benjamini and Hochberg method and a FDR cut-off of 5% to threshold significance.

After removing chimeric results, our analysis of the Cufflinks RABT transcriptome predicted 3815 differentially expressed genes, 127 alternative splicing events, and 81 cases of alternative promoter usage [**see GSE57439**]. This included 354 significant intergenic lncRNA loci (out of 647) and 69 significant antisense loci (out of 467). We noted an absence of strong correlation between maximum sense and antisense partner expression [$\rho = .19$, **Supplementary Figure 3.12**], which reiterated the strand-specificity of our sequencing reads and the integrity of the antisense signal. Analyzing the *PlasmoDBv10.0* annotated transcriptome alone predicted 4284 differentially expressed mRNA genes.

GO Term Enrichment

We defined 1632 ring-specific genes as differentially expressed, annotated genes with maximum expression between 6-20 hpi during the 56-hour time course. Similarly, we defined 1378 trophozoite-specific genes (24-36 hpi maximum expression) and 1274 schizont-specific genes (40-48 hpi maximum expression). We converted *PlasmoDBv10.0* gene annotations to SangerDB *P. falciparum* gene annotations, and then used GOstat and a FDR cut-off of 1% to search for biological process GO terms significantly over-represented in the ring-specific, trophozoite-specific, and schizont-specific gene sets [70]. See **Supplementary Table 3.3** for stage-specific results.

We intersected the 1381 predicted circRNA loci with *PlasmoDBv10.0* gene annotations to yield a circRNA-producing gene set, and used an analogous approach to search for

significantly over-represented GO terms. In the circRNA GO term analysis, we included biological process, molecular function, and cellular component results [**Supplementary Table 3.6**].

Sample Staging and Expression Profile Visualizations

To visualize sample similarities on a transcriptome-wide scale, we first constructed a distance matrix of sample expression profiles as follows: (1) we only considered contributions from the 5077 annotated mRNAs with an FPKM rising above one during the 56-hour time course, (2) we transformed FPKM values as log2(FPKM+1), (3) we computed the Pearson correlation (ρ) matrix between sample expression profiles, and (4) we transformed this matrix as (1- ρ)/2. We then used MDS as implemented by the R-project *cmdscale* function to visualize sample expression profile similarities in two dimensions.

To visualize differentially expressed, annotated mRNA expression profile dynamics during the 56-hour time course, we followed steps 1-2 as above using the 4284 differentially expressed mRNA genes, and then used the *nMDS* and *circularmap* functions distributed in the R-project NeatMap package [68]. The NeatMap package implements the non-metric MDS algorithm previously proposed by Taguchi and Oono [68,104]. We used the more traditional *make.heatmap1* function distributed in the R-project NeatMap package to compare the expression profiles of lncRNAs and annotated mRNAs (FPKM >0) during the 56-hour time course.

Stage Specificity Score

To summarize transcript stage specificity in a single quantitative metric, we calculated a maximum transcript specificity score that is based on the Jensen-Shannon (JS) divergence. This

metric has been fully described by Cabili et al. and was used by this group to annotate the tissue specificity of human intergenic lncRNA transcripts [76]. Specifically, we calculated specificity scores for each transcript, inspected the top 5% most stage-specific transcripts, and compared transcript specificity scores by transcript class. We normalized transcript expression levels as in Cabili et al. and used the *shannon.entropy, JSdistVec,* and *JSdistFromP* functions implemented in the R-project CummeRbund package, except that we used a logarithmic base of two in the Shannon entropy calculation to match the Cabili et al. methods [71,76]. We also performed this analysis on the set of 1619 annotated mRNAs with an average FPKM of less than 20 (set metrics: average average FPKM 1.95; average maximum FPKM 4.3).

Corroborating our visual observations, each telomere-associated lncRNA-TARE transcript was in the set of top stage-specific transcripts [**see GSE57439**]. However, global comparisons of lncRNA specificity scores and mRNA specificity scores were less conclusive; lncRNAs appeared to be more stage-specific than the full set of annotated mRNAs, but this difference was not clear when we compared lncRNAs to only lowly expressed mRNAs [**Supplementary Figure 3.13**]. Thus, while the specificity metric accurately summarized the increased stage specificity of certain transcripts, such as the telomere-associated lncRNA-TAREs, further comparisons will be necessary to make any global conclusions as to the increased or equivalent specificity of *P. falciparum* lncRNAs versus mRNAs.

Expression Correlation

To generate a null distribution of location-independent gene pair correlation, we investigated the Pearson correlation during the 56-hour time course between 50,000 random *PlasmoDBv10.0* mRNA gene pairs. We then computed the Pearson correlation between 5251 mRNA-neighboring gene pairs, 498 intergenic lncRNA-neighboring gene pairs, and 445 sense-

antisense gene pairs, and compared these distributions to the null distribution of location-independent gene pair correlation. To be consistent, we defined the neighboring gene of both mRNAs and intergenic lncRNAs as their more correlated neighboring mRNA (unless otherwise specified). We excluded gene pairs from analysis when one of the genes was a short and/or structural RNA, or otherwise had an FPKM of zero during the 56-hour time course. Also, we excluded the result if the more or less correlated neighbor of an intergenic lncRNA was another intergenic lncRNA.

Using this approach, we found a significant shift towards positive correlation in the case of both mRNA-neighboring genes and intergenic lncRNA-neighboring genes (Wilcoxon rank sum p-value < 2.2e-16 in both cases), and a significant shift towards negative correlation in the case of sense-antisense genes (Wilcoxon rank sum p-value = 3.834e-11). Our results also indicated that lncRNA-neighboring gene pairs were significantly more highly correlated than mRNA-neighboring gene pairs (Wilcoxon rank sum p-value < 2.2e-16). We pursued this result further. See **Supplementary Figure 3.6** for details.

LncRNA Length, GC content, and Antisense Relative Location

We only considered exonic sequence in lncRNA transcript length and GC content calculations. However, we included both intronic and exonic sequence in the antisense transcript location and annotated gene body length normalizations.

CircRNA Prediction

We predicted 1381 *P. falciparum* circRNAs using the methods developed by Memczak et al. [83]. In brief, we aligned reads to the *PlasmoDBv10.0* genome using Bowtie v2.1.0, and used circBase v1.0 scripts to split high-quality unaligned reads into anchors, subsequently screening

anchors for linear or head-to-tail (circular) splicing. To get a reasonable set of circRNA candidates, we used the default circBase filtering suggestions, except for specifying a minimum circRNA size of 100 bp and a maximum circRNA size of 10,000 bp.

CircRNA Primer Design

We designed convergent and divergent PCR primers for nine *P. falciparum* genes with predicted circRNAs using AlleleID v7.6 software (Premier Biosoft International). Furthermore, we designed divergent forward primers as the reverse complement of convergent reverse primers, and divergent reverse primers as the reverse complement of convergent forward primers. We then added a tail consisting of the T7 promoter sequence 5'-TAA TAC GAC TCA CTA TAG GG-3' and 5'-TAG TAG TAG TAG-3' to all forward primers, and a tail consisting of the M13F(-47) sequence 5'-CGC CAG GGT TTT CCC AGT CAC GAC-3' and 5'-TAG TAG TAG TAG-3' to all reverse primers. This allowed us to directly sequence both divergent and convergent PCR amplicons using universal T7 or M13F(-47) primers. Primer and expected amplicon sequences are listed in **Supplementary Table 3.7**.

3.5 References

1. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, et al. (2014) A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature 505: 50-55.
2. Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, et al. (2010) Origin of the human malaria parasite Plasmodium falciparum in gorillas. Nature 467: 420-425.
3. Males S, Gaye O, Garcia A (2008) Long-term asymptomatic carriage of Plasmodium falciparum protects from malaria attacks: a prospective study among Senegalese children. Clin Infect Dis 46: 516-522.
4. Mita T, Tanabe K (2012) Evolution of Plasmodium falciparum drug resistance: implications for the development and containment of artemisinin resistance. Jpn J Infect Dis 65: 465-475.
5. Murray CJ, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, et al. (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet 379: 413-431.
6. Goldberg DE, Cowman AF (2010) Moving in and renovating: exporting proteins from Plasmodium into host erythrocytes. Nature Reviews Microbiology 8: 617-621.
7. Maier AG, Cooke BM, Cowman AF, Tilley L (2009) Malaria parasite proteins that remodel the host erythrocyte. Nature Reviews Microbiology 7: 341-354.
8. Bannister LH, Hopkins JM, Fowler RE, Krishna S, Mitchell GH (2000) A brief illustrated guide to the ultrastructure of Plasmodium falciparum asexual blood stages. Parasitol Today 16: 427-433.
9. Miller LH, Baruch DI, Marsh K, Doumbo OK (2002) The pathogenic basis of malaria. Nature 415: 673-679.
10. Baker DA (2010) Malaria gametocytogenesis. Mol Biochem Parasitol 172: 57-65.
11. Trager W, Jensen JB (1976) Human Malaria Parasites in Continuous Culture. Science 193: 673-675.
12. Burrows JN, Burlot E, Campo B, Cherbuin S, Jeanneret S, et al. (2014) Antimalarial drug discovery - the path towards eradication. Parasitology 141: 128-139.
13. Sinden RE, Carter R, Drakeley C, Leroy D (2012) The biology of sexual development of Plasmodium: the design and implementation of transmission-blocking strategies. Malar J 11: 70.
14. Miller LH, Ackerman HC, Su XZ, Wellems TE (2013) Malaria biology and disease pathogenesis: insights for new treatments. Nat Med 19: 156-167.
15. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.
16. Garcia CR, de Azevedo MF, Wunderlich G, Budu A, Young JA, et al. (2008) Plasmodium in the postgenomic era: new insights into the molecular cell biology of malaria parasites. Int Rev Cell Mol Biol 266: 85-156.
17. Winzeler EA (2008) Malaria research in the post-genomic era. Nature 455: 751-756.
18. Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu JC, et al. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. Plos Biology 1: 85-100.
19. Le Roch KG, Zhou YY, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503-1508.
20. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, et al. (2013) Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum. Genome Biol 14: R128.
21. Foth BJ, Zhang N, Mok S, Preiser PR, Bozdech Z (2008) Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites. Genome Biol 9: R177.
22. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, et al. (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. Genome Res 14: 2308-2318.
23. Cui L, Miao J (2010) Chromatin-mediated Epigenetic Regulation in the Malaria Parasite Plasmodium falciparum. Eukaryot Cell.
24. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, et al. (2010) H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. Plos Pathogens 6: e1001223.
25. Chaal BK, Gupta AP, Wastuwidyaningtyas BD, Luah YH, Bozdech Z (2010) Histone Deacetylases Play a Major Role in the Transcriptional Regulation of the Plasmodium falciparum Life Cycle. Plos Pathogens 6: -.

26. Gupta AP, Chin WH, Zhu L, Mok S, Luah YH, et al. (2013) Dynamic epigenetic regulation of gene expression during the life cycle of malaria parasite Plasmodium falciparum. PLoS Pathog 9: e1003170.
27. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, et al. (2012) Transcriptional variation in the malaria parasite Plasmodium falciparum. Genome Res 22: 925-938.
28. Comeaux CA, Coleman BI, Bei AK, Whitehurst N, Duraisingh MT (2011) Functional analysis of epigenetic regulation of tandem RhopH1/clag genes reveals a role in Plasmodium falciparum growth. Molecular Microbiology 80: 378-390.
29. Cortes A, Carret C, Kaneko O, Yim Lim BY, Ivens A, et al. (2007) Epigenetic silencing of Plasmodium falciparum genes linked to erythrocyte invasion. PLoS Pathog 3: e107.
30. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, et al. (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. Cell 121: 25-36.
31. Chookajorn T, Dzikowski R, Frank M, Li F, Jiwani AZ, et al. (2007) Epigenetic memory at malaria virulence genes. Proc Natl Acad Sci U S A 104: 899-902.
32. Dzikowski R, Frank M, Deitsch K (2006) Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. PLoS Pathog 2: e22.
33. Eksi S, Morahan BJ, Haile Y, Furuya T, Jiang H, et al. (2012) Plasmodium falciparum gametocyte development 1 (Pfgdv1) and gametocytogenesis early gene identification and commitment to sexual development. PLoS Pathog 8: e1002964.
34. Wright GJ, Rayner JC (2014) Plasmodium falciparum erythrocyte invasion: combining function with immune evasion. PLoS Pathog 10: e1003943.
35. Yap A, Azevedo MF, Gilson PR, Weiss GE, O'Neill MT, et al. (2014) Conditional expression of apical membrane antigen 1 in Plasmodium falciparum shows it is required for erythrocyte invasion by merozoites. Cell Microbiol 16: 642-656.
36. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, et al. (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res.
37. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, et al. (2014) A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature 507: 248-252.
38. Guizetti J, Scherf A (2013) Silence, activate, poise and switch! Mechanisms of antigenic variation in Plasmodium falciparum. Cell Microbiol 15: 718-726.
39. Balaji S, Babu MM, Iyer LM, Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res 33: 3994-4006.
40. Coulson RM, Hall N, Ouzounis CA (2004) Comparative genomics of transcriptional control in the human malaria parasite Plasmodium falciparum. Genome Res 14: 1548-1554.
41. Baum J, Papenfuss AT, Mair GR, Janse CJ, Vlachou D, et al. (2009) Molecular genetics and comparative genomics reveal RNAi is not functional in malaria parasites. Nucleic Acids Res 37: 3788-3798.
42. Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, et al. (2011) A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. Genome biology 12: R56.
43. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, et al. (2011) Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. Bmc Genomics 12: 587.
44. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, et al. (2010) New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. Molecular Microbiology 76: 12-24.
45. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, et al. (2014) Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. BMC Genomics 15: 150.
46. Wang PL, Bao Y, Yee MC, Barrett SP, Hogan GJ, et al. (2014) Circular RNA is expressed across the eukaryotic tree of life. PLoS One 9: e90859.
47. Sorber K, Dimon MT, DeRisi JL (2011) RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. Nucleic Acids Res 39: 3820-3835.
48. Lee JT (2012) Epigenetic regulation by long noncoding RNAs. Science 338: 1435-1439.

49. Pelechano V, Steinmetz LM (2013) Gene regulation by antisense transcription. Nat Rev Genet 14: 880-893.
50. Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76-79.
51. Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N (1996) Requirement for Xist in X chromosome inactivation. Nature 379: 131-137.
52. van Werven FJ, Neuert G, Hendrick N, Lardenois A, Buratowski S, et al. (2012) Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. Cell 150: 1170-1181.
53. Hongay CF, Grisafi PL, Galitski T, Fink GR (2006) Antisense transcription controls cell fate in Saccharomyces cerevisiae. Cell 127: 735-745.
54. Swiezewski S, Liu F, Magusin A, Dean C (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature 462: 799-802.
55. Pontier DB, Gribnau J (2011) Xist regulation and function explored. Hum Genet 130: 223-236.
56. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology 12: R18.
57. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, et al. (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods 7: 709-715.
58. Lopez-Barragan MJ, Quinones M, Cui K, Lemieux J, Zhao K, et al. (2011) Effect of PCR extension temperature on high-throughput sequencing. Mol Biochem Parasitol 176: 64-67.
59. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, et al. (2012) Optimal enzymes for amplifying sequencing libraries. Nature methods 9: 10-11.
60. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37: e123.
61. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, et al. (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods 10: 623-629.
62. Su XZ, Wu Y, Sifri CD, Wellems TE (1996) Reduced extension temperatures required for PCR amplification of extremely A+T-rich DNA. Nucleic Acids Res 24: 1574-1575.
63. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.
64. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, et al. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics 28: 1530-1532.
65. Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28: 2184-2185.
66. Lemieux JE, Gomez-Escobar N, Feller A, Carret C, Amambua-Ngwa A, et al. (2009) Statistical estimation of cell-cycle progression and lineage commitment in Plasmodium falciparum reveals a homogeneous pattern of transcription in ex vivo culture. Proc Natl Acad Sci U S A 106: 7559-7564.
67. Joice R, Narasimhan V, Montgomery J, Sidhu AB, Oh K, et al. (2013) Inferring developmental stage composition from gene expression in human malaria. PLoS Comput Biol 9: e1003392.
68. Rajaram S, Oono Y (2010) NeatMap--non-clustering heat map alternatives in R. BMC Bioinformatics 11: 45.
69. Tzeng J, Lu HH, Li WH (2008) Multidimensional scaling for large genomic data sets. BMC Bioinformatics 9: 179.
70. Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20: 1464-1465.
71. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 7: 562-578.
72. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8: 1494-1512.
73. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515.
74. Lu B, Zeng Z, Shi T (2013) Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci China Life Sci 56: 143-155.
75. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12: 671-682.

76. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25: 1915-1927.
77. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, et al. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505: 635-640.
78. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46-58.
79. Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F (2007) Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae. Cell 131: 706-717.
80. Lapidot M, Pilpel Y (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep 7: 1216-1222.
81. Epp C, Li F, Howitt CA, Chookajorn T, Deitsch KW (2009) Chromatin associated sense and antisense noncoding RNAs are transcribed from the var gene family of virulence genes of the malaria parasite Plasmodium falciparum. RNA 15: 116-127.
82. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, et al. (2013) Natural RNA circles function as efficient microRNA sponges. Nature 495: 384-388.
83. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, et al. (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 495: 333-338.
84. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, et al. (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 19: 141-157.
85. Wang L, Jiang N, Wang L, Fang O, Leach LJ, et al. (2014) 3' Untranslated regions mediate transcriptional interference between convergent genes both locally and ectopically in Saccharomyces cerevisiae. PLoS Genet 10: e1004021.
86. Gullerova M, Proudfoot NJ (2010) Transcriptional interference and gene orientation in yeast: noncoding RNA connections. Cold Spring Harb Symp Quant Biol 75: 299-311.
87. Castelnuovo M, Rahman S, Guffanti E, Infantino V, Stutz F, et al. (2013) Bimodal expression of PHO84 is modulated by early termination of antisense transcription. Nat Struct Mol Biol 20: 851-858.
88. Modarresi F, Faghihi MA, Lopez-Toledano MA, Fatemi RP, Magistri M, et al. (2012) Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. Nat Biotechnol 30: 453-459.
89. Morris KV, Santoso S, Turner AM, Pastori C, Hawkins PG (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. PLoS Genet 4: e1000258.
90. Magistri M, Faghihi MA, St Laurent G, 3rd, Wahlestedt C (2012) Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. Trends Genet 28: 389-396.
91. Kolevzon N, Nasereddin A, Naik S, Yavin E, Dzikowski R (2014) Use of peptide nucleic acids to manipulate gene expression in the malaria parasite Plasmodium falciparum. PLoS One 9: e86802.
92. Barker RH, Jr., Metelev V, Rapaport E, Zamecnik P (1996) Inhibition of Plasmodium falciparum malaria using antisense oligodeoxynucleotides. Proc Natl Acad Sci U S A 93: 514-518.
93. Rapaport E, Misiura K, Agrawal S, Zamecnik P (1992) Antimalarial activities of oligodeoxynucleotide phosphorothioates in chloroquine-resistant Plasmodium falciparum. Proc Natl Acad Sci U S A 89: 8577-8580.
94. Bright AT, Winzeler EA (2011) Noncoding RNA, antigenic variation, and the virulence genes of Plasmodium falciparum. BMC Biol 9: 50.
95. Deitsch KW, Calderwood MS, Wellems TE (2001) Malaria. Cooperative silencing elements in var genes. Nature 412: 875-876.
96. Kirchmaier AL, Rine J (2001) DNA replication-independent silencing in S. cerevisiae. Science 291: 646-650.
97. Li YC, Cheng TH, Gartenberg MR (2001) Establishment of transcriptional silencing in the absence of DNA replication. Science 291: 650-653.
98. LaMonte G, Philip N, Reardon J, Lacsina JR, Majoros W, et al. (2012) Translocation of sickle cell erythrocyte microRNAs into Plasmodium falciparum inhibits parasite translation and contributes to malaria resistance. Cell Host Microbe 12: 187-199.
99. Lambros C, Vanderberg JP (1979) Synchronization of Plasmodium falciparum erythrocytic stages in culture. J Parasitol 65: 418-420.

100. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210.
101. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al. (2006) GenePattern 2.0. Nat Genet 38: 500-501.
102. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
103. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841-842.
104. Taguchi YH, Oono Y (2005) Relational patterns of gene expression via non-metric multidimensional scaling analysis. Bioinformatics 21: 730-740.

Chapter 4 | **Discussion**

4.1 Concluding remarks

Pervasive transcription of lncRNA has been described for nearly every eukaryotic genome [1]. However, albeit clear examples of functional lncRNAs in humans and model organisms, it is widely accepted that not all lncRNAs are functionally important [2-4]. Thus, it is necessary to systematically identify, classify, and prioritize lncRNA candidates for functional investigation [3,5,6].

To identify lncRNAs, microarray-based or RNA sequencing approaches have been classically used, and in this thesis we apply both technologies to *P. falciparum* transcriptome analysis. In **Chapter 2**, we use a high-resolution DNA tiling array to annotate 60 non-coding, intergenic, transcriptionally active regions (TARs), and characterize a compelling family of telomere-associated lncRNAs [7]. In **Chapter 3**, we develop and apply strand-specific, non-polyA-selected RNA sequencing. Using this approach, we pursue the first *P. falciparum* transcriptome assembly from RNA sequencing reads, and assemble 660 intergenic lncRNA and 474 antisense RNA transcript structures. We also investigate and predict 1381 non-canonical circular RNA splice junctions.

In terms of lncRNA classification, some natural classifications exist, and may be generally predictive of function [5]. For example, lncRNAs can be grouped by their relationship to known transcripts and regulatory elements: antisense, intronic, intergenic, circular, pseudogenic, promoter-associated, or enhancer-associated. In **Chapter 2** and **Chapter 3**, we extensively analyze *P. falciparum* lncRNA transcript properties, including lncRNA relationship

to known transcripts, expression, stage-specificity, splicing, length, GC content, and conservation. Our analyses reveal global trends for intergenic lncRNAs versus antisense RNAs, as well as outliers that may represent specific subclasses. As an example of how global analysis can predict function, we find that lncRNAs are developmentally regulated, and that sense-antisense transcript pairs are often inversely expressed. This suggests that much of *P. falciparum* lncRNA may play a regulatory role, especially via antagonistic mechanisms in the case of antisense RNAs.

The end goal of this thesis work was to prioritize lncRNA candidates for functional investigation. Our top candidates include a family of twenty-two telomere-associated lncRNAs whose expression peaks during parasite invasion, as well as a spliced, highly correlated antisense counterpart to an essential gene in the early gametocyte commitment program. We also find additional examples of highly expressed, developmentally regulated, spliced lncRNAs, and an intriguing subset of antisense RNAs whose expression markedly drops during parasite invasion.

Taken together, the work herein confirms the regulatory capacity of *P. falciparum* lncRNA. Our results provide an immediate foundation for future in depth functional analyses of *P. falciparum* lncRNAs. Our methods and analysis pipeline also enable further genome-wide studies of lncRNAs in sexual blood and mosquito stages or during growth perturbations.

4.2 Future directions

The CRISPR (clustered regularly interspaced short palindromic repeats) system functions as an immune system in many prokaryotic organisms [8]. In this natural system, when foreign DNA is introduced, a unique spacer is created and integrated into the genome. Transcribed CRISPR spacers, in complex with a CRISPR-associated (Cas) nuclease, then cut and inactivate

foreign DNA [8-10]. In early 2013, Mali et al. and Cong et al. ported the CRISPR system to human cell culture, and demonstrated its utility in human genome editing [11,12]. Also in 2013, the CRISPR genome editing technique was applied to numerous model organisms: mice [13], zebra fish [14], yeast [15], nematodes [16], and plants [17], amongst others. Importantly, the CRISPR system is more specific than RNA interference-based approaches, and the effect is permanent [18-21]. The CRISPR system is also far less cumbersome, expensive, and time-consuming than previous genome editing techniques, as the specificity for each target is imparted through the design of a custom guide RNA rather than a custom protein [10].

Experimental investigation of *P. falciparum* lncRNA should be dramatically more feasible given the recent adaptation of the CRISPR genome editing system to *P. falciparum* [22]. Ghorbal et al. have demonstrated that CRISPR/Cas9 can be used to replace an endogenous *P. falciparum* target gene with a drug-selectable marker. More significantly, this group has used CRISPR/Cas9 to introduce single-nucleotide base changes into targeted loci in the *P. falciparum* genome without integrating a drug-selectable marker [22]. *P. falciparum* genetic engineering has been notoriously intractable for decades, but the flexibility and efficiency of the CRISPR system is revolutionary. Taken together, this timely technical advancement greatly expands the *P. falciparum* genetic engineering toolkit, such that investigating *P. falciparum* lncRNA function and mechanism using genetic dissections routinely performed in model organisms is now tractable. A suggested first set of experiments is: (1) to edit endogenous lncRNA promoter sequences such that transcription is abrogated, temporally offset, and/or constitutive, and (2) to assay for putative phenotypic consequences.

Additional systematic studies of *P. falciparum* lncRNAs to generate further hypotheses are called for as well. Strand-specific deep profiling of the transition from asexual to sexual

development, gametocyte maturation, and mosquito stages is fundamentally needed. In model organisms, profling lncRNAs in mutants lacking a component of the exosome has been particularly revealing [23-26], as has profiling the lncRNA transcriptional response to heat shock and various other environmental perturbations [27]. *P. falciparum* possesses conserved exosome components which could similarly be knocked out to investigate roles for unstable non-coding transcription [28]. Perturbation assays to investigate changes in *P. falciparum* lncRNA expression caused by fluctuating nutrient and/or oxygen levels, and febrile temperature would also be important to explore, as there is a clear precedent for lncRNA-mediated integration of various environmental signals in other organisms [27,29-33].

To date, most *P. falciparum* genome-wide studies have focused on one or two aspects of suspected gene regulation, such as steady-state transcript levels, RNA decay rate, chromatin state, or protein levels [28,34-37]. Albeit a massive undertaking, pairing strand-specific deep sequencing with polysome-associated RNA profiling, ChIP-sequencing, chromatin capture, and/or peptide profiling in matched samples would provide a much more cohesive picture of *P. falciparum* gene regulation. Moreover, most *P. falciparum* genome-wide studies have only incorporated a handful of samples from one blood stage time-course. Our results, however, demonstrate the added benefit of incorporating biological replication, sampling parasites at approximately four-hour intervals, and assaying at least 56 hours of blood stage development.

A final comment is that the analyses and interpretations presented throughout this thesis are based on populations of cells. However, the story may be considerably different within single cells. Indeed, the few genome-wide studies that have investigated single cell transcriptional dynamics have documented considerable heterogeneity [38]. Adaption of such single-cell approaches to the AT rich *P. falciparum* transcriptome is likely on the horizon [39]. Immediately

feasible, albeit lower throughput, is fluorescence in situ hybridization (FISH). FISH can be used to quantify transcript abundance for a handful of mRNA and/or lncRNA targets in single cells [40]. FISH also reveals the localization compartment of lncRNAs at a snapshot in time, which provides an important clue to their functional role [41 ,42].

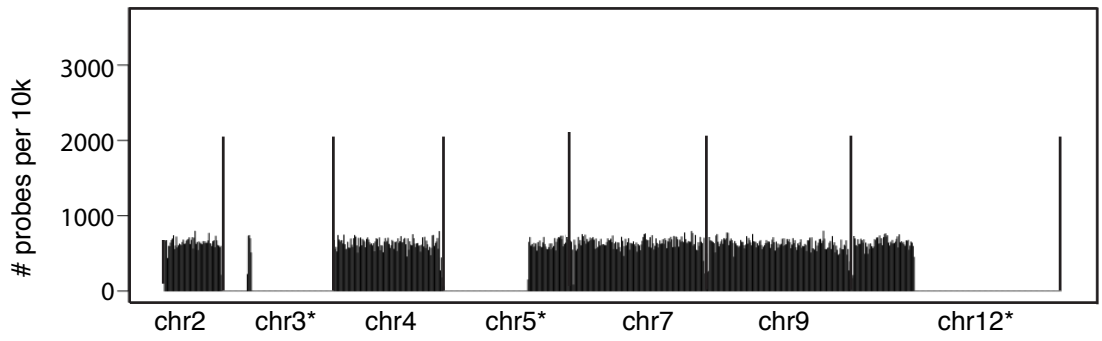
4.3 References

1. Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nature reviews Genetics 10: 833-844.
2. Kowalczyk MS, Higgs DR, Gingeras TR (2012) Molecular biology: RNA discrimination. Nature 482: 310-311.
3. Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. Cell 154: 26-46.
4. Baker M (2011) Long noncoding RNAs: the search for function. Nature Methods 8: 379-383.
5. Ma L, Bajic VB, Zhang Z (2013) On the classification of long non-coding RNAs. RNA biology 10: 925-933.
6. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25: 1915-1927.
7. Broadbent KM, Park D, Wolf AR, Van Tyne D, Sims JS, et al. (2011) A global transcriptional analysis of Plasmodium falciparum malaria reveals a novel family of telomere-associated lncRNAs. Genome biology 12: R56.
8. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709-1712.
9. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, et al. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816-821.
10. Pennisi E (2013) The CRISPR craze. Science (New York, N Y) 341: 833-836.
11. Mali P, Yang L, Esvelt KM, Aach J, Guell M, et al. (2013) RNA-guided human genome engineering via Cas9. Science (New York, N Y) 339: 823-826.
12. Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. Science (New York, N Y) 339: 819-823.
13. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, et al. (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153: 910-918.
14. Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, et al. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature biotechnology 31: 227-229.
15. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, et al. (2013) Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic acids research 41: 4336-4343.
16. Friedland AE, Tzur YB, Esvelt KM, Colaiacovo MP, Church GM, et al. (2013) Heritable genome editing in C. elegans via a CRISPR-Cas9 system. Nature methods 10: 741-743.
17. Jiang W, Zhou H, Bi H, Fromm M, Yang B, et al. (2013) Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. Nucleic acids research 41: e188.
18. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, et al. (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154: 442-451.
19. Heintze J, Luft C, Ketteler R (2013) A CRISPR CASe for high-throughput silencing. Frontiers in genetics 4: 193.
20. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, et al. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152: 1173-1183.
21. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nature biotechnology 32: 279-284.
22. Ghorbal M, Gorman M, Macpherson CR, Martins RM, Scherf A, et al. (2014) Genome editing in the human malaria parasite Plasmodium falciparum using the CRISPR-Cas9 system. Nat Biotechnol.
23. Davis CA, Ares M, Jr. (2006) Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 103: 3262-3267.
24. Camblong J, Iglesias N, Fickentscher C, Dieppois G, Stutz F (2007) Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae. Cell 131: 706-717.
25. Chekanova JA, Gregory BD, Reverdatto SV, Chen H, Kumar R, et al. (2007) Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the Arabidopsis transcriptome. Cell 131: 1340-1353.
26. Buhler M, Spies N, Bartel DP, Moazed D (2008) TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway. Nature structural & molecular biology 15: 1015-1023.

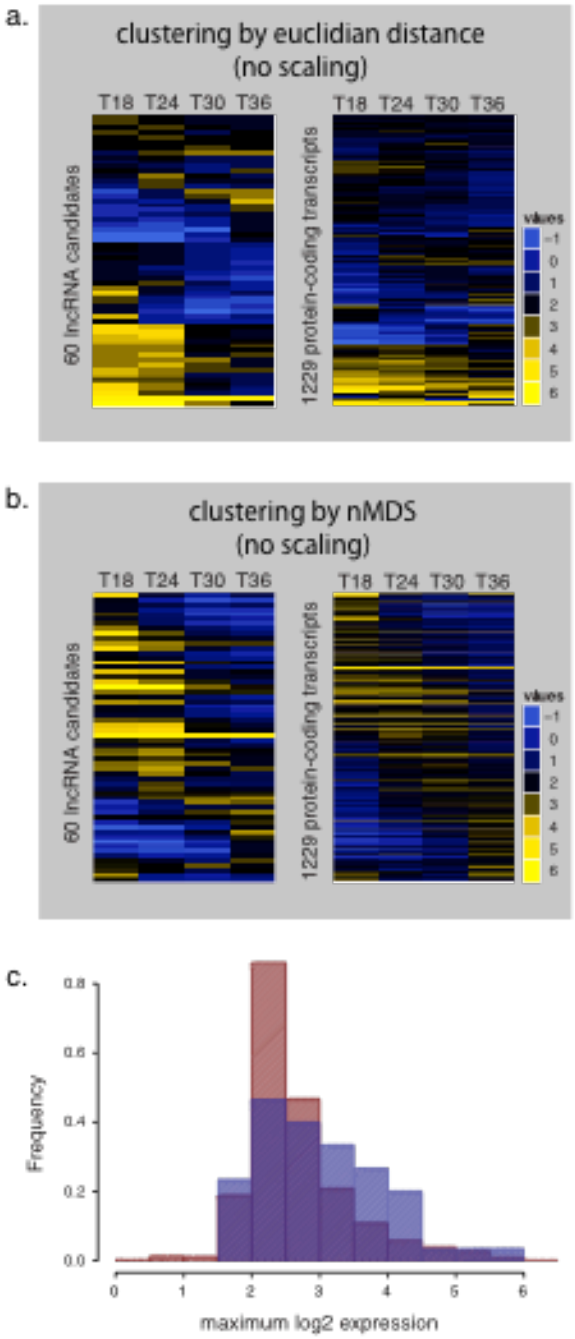
27. Lakhotia SC (2012) Long non-coding RNAs coordinate cellular responses to stress. Wiley interdisciplinary reviews RNA 3: 779-796.
28. Shock JL, Fischer KF, DeRisi JL (2007) Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. Genome biology 8: R134.
29. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311-1323.
30. Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331: 76-79.
31. Houseley J, Rubbi L, Grunstein M, Tollervey D, Vogelauer M (2008) A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster. Mol Cell 32: 685-695.
32. Hainer SJ, Pruneski JA, Mitchell RD, Monteverde RM, Martens JA (2011) Intergenic transcription causes repression by directing nucleosome assembly. Genes Dev 25: 29-40.
33. van Werven FJ, Amon A (2011) Regulation of entry into gametogenesis. Philosophical transactions of the Royal Society of London Series B, Biological sciences 366: 3521-3531.
34. Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, et al. (2010) H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. Plos Pathogens 6: e1001223.
35. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, et al. (2013) Polysome profiling reveals translational control of gene expression in the human malaria parasite Plasmodium falciparum. Genome Biol 14: R128.
36. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, et al. (2011) Directional gene expression and antisense transcripts in sexual and asexual stages of Plasmodium falciparum. Bmc Genomics 12: 587.
37. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, et al. (2014) Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. BMC Genomics 15: 150.
38. Huang S (2009) Non-genetic heterogeneity of cells in development: more than just noise. Development (Cambridge, England) 136: 3853-3862.
39. Macaulay IC, Voet T (2014) Single cell genomics: advances and future perspectives. PLoS genetics 10: e1004126.
40. Itzkovitz S, van Oudenaarden A (2011) Validating transcripts with probes and imaging technology. Nature methods 8: S12-19.
41. Castelnuovo M, Rahman S, Guffanti E, Infantino V, Stutz F, et al. (2013) Bimodal expression of PHO84 is modulated by early termination of antisense transcription. Nat Struct Mol Biol 20: 851-858.
42. Chakraborty D, Kappei D, Theis M, Nitzsche A, Ding L, et al. (2012) Combined RNAi and localization for functionally dissecting long noncoding RNAs. Nature methods 9: 360-362.

Appendix

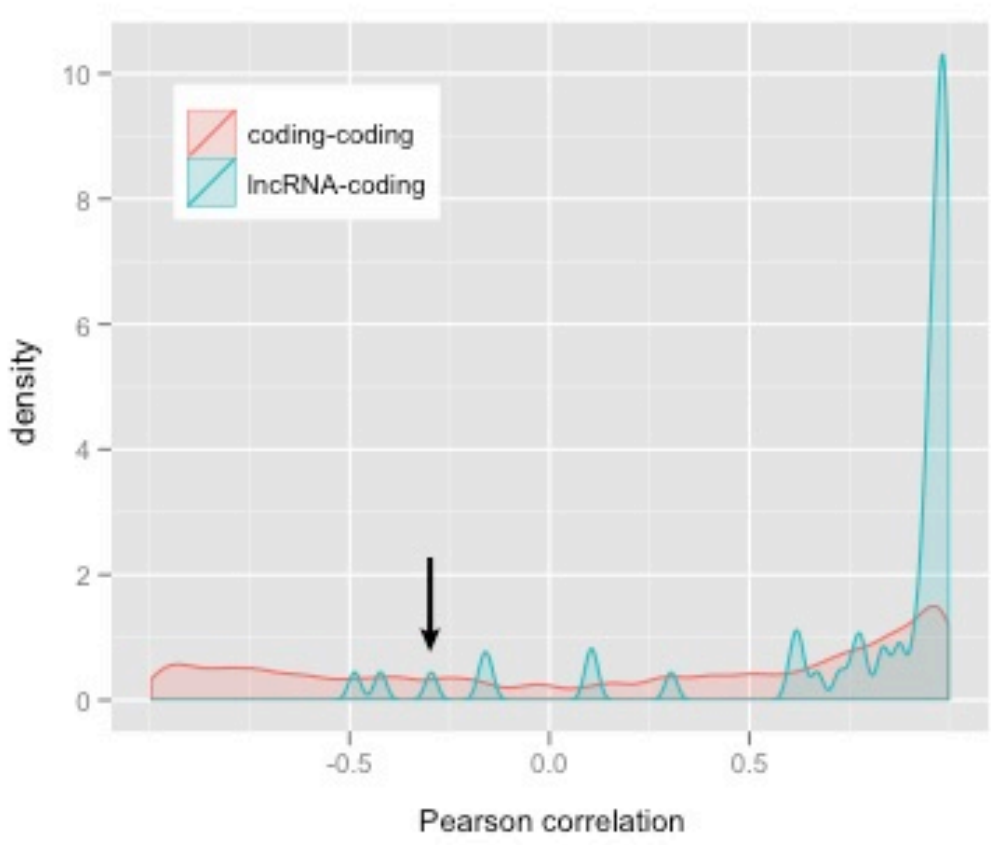
Chapter 2 Supplementary Figures and Tables



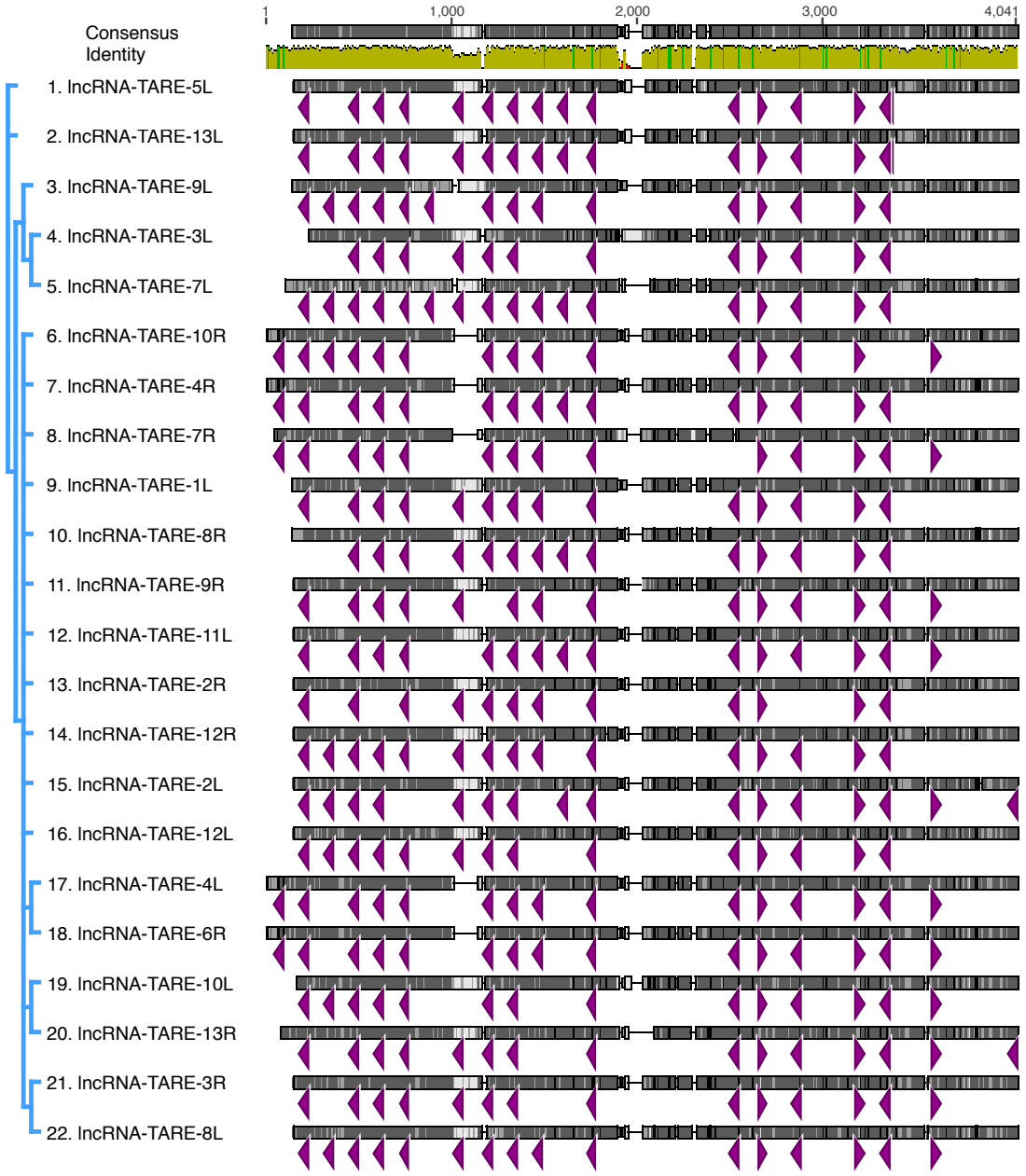
Supplementary Figure 2.1 | DNA tiling array genome coverage is dense and systematic. *P. falciparum* strain 3D7 chromosomes 2, 4, 7, and 9 are completely tiled. Chromosomes 3, 5, and 12 are partially tiled (*). Number of probes per 10,000 bp is plotted versus chromosomal position. The median probe resolution is 12 bp, corresponding to approximately 830 probes per 10,000 bp.



Supplementary Figure 2.2 | LncRNAs and protein-coding transcripts display similar expression. (a) Hierarchical clustering of lncRNAs and protein-coding transcripts by euclidian distance without mean centering across time-points. **(b)** Clustering by non-Metric Multidimensional Scaling (nMDS) without mean centering across time-points. **(c)** Histogram of maximum log2 expression in any time-point for lncRNAs (blue) and protein-coding transcripts (red).

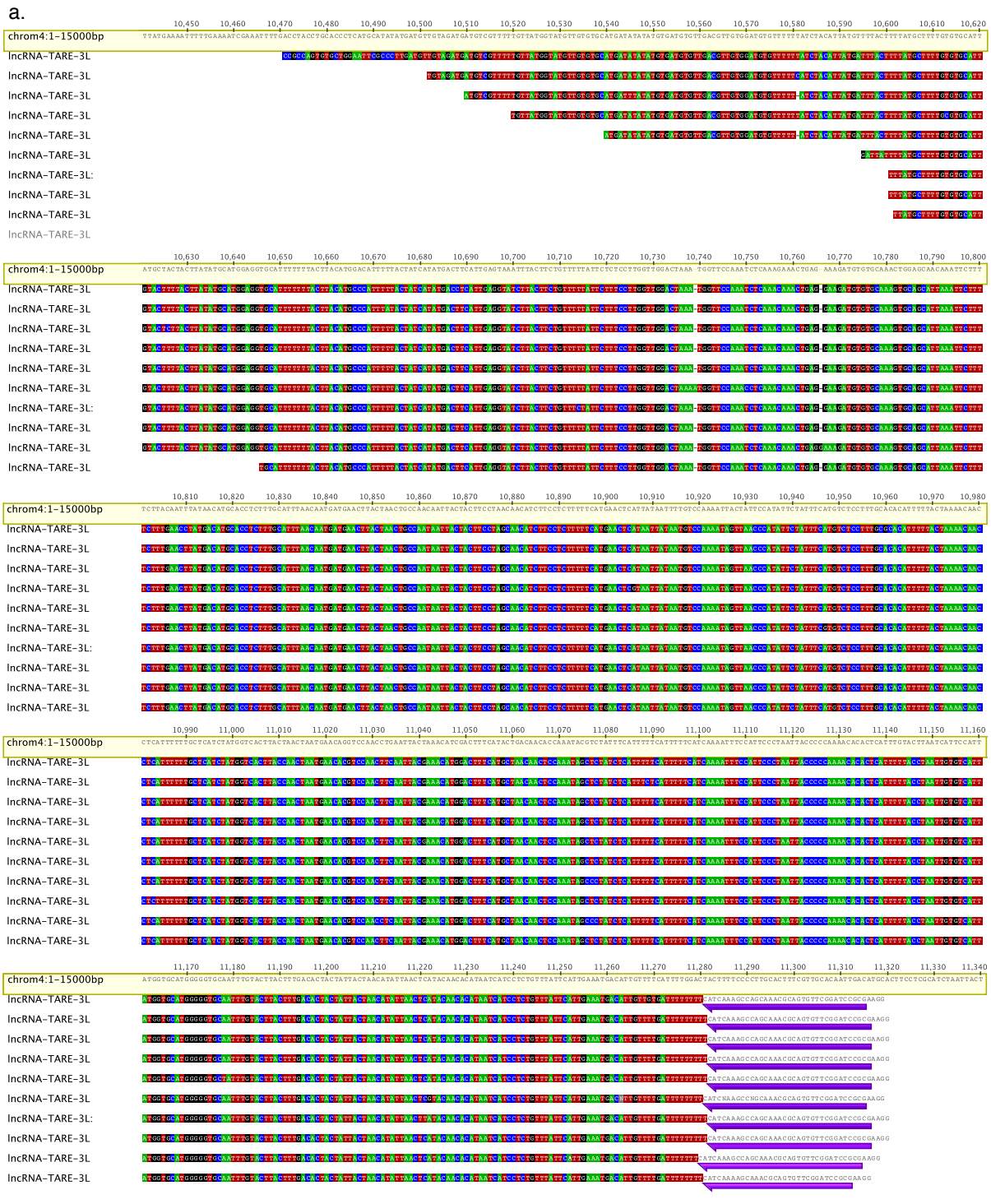


Supplementary Figure 2.3 | Distribution of time course correlations between the 60 putative lncRNAs and their neighboring coding genes (teal) versus correlations between coding genes and their neighboring coding genes (red). The set of 60 putative lncRNAs is enriched for transcripts that are highly correlated to nearby genes, suggesting that some are spliced or un-annotated untranslated regions (UTRs) to genes. The arrow marks the correlation of lncRNA-TARE-4L to its neighbor.



Supplementary Figure 2.4 | LncRNA-TARE homology. Clustering, conservation, and SPE2 motifs within the 22 predicted lncRNA-TARE loci. Figure produced using Geneious software. Clustering performed using ClustalW.

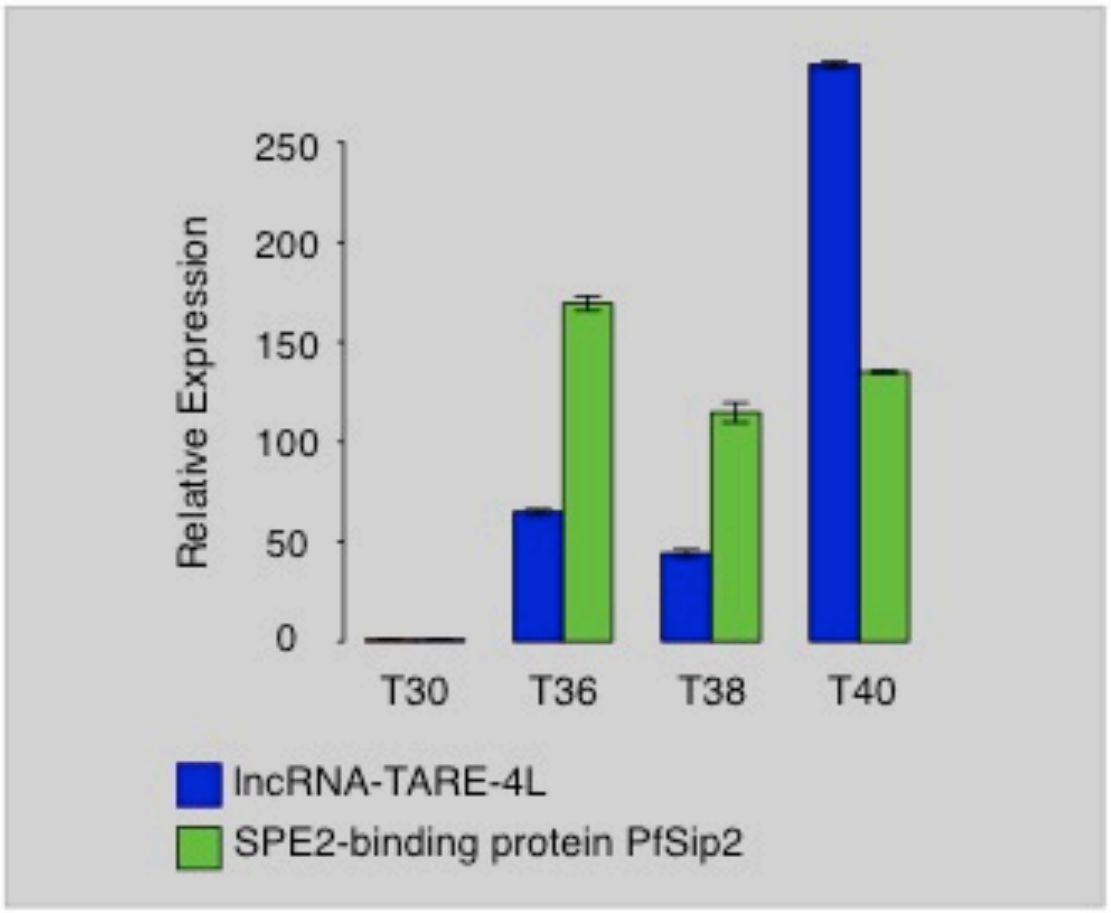
Supplementary Figure 2.5 | Sequenced lncRNA-TARE RLM-RACE products. Products were BLASTed, trimmed to exclude excessive vector sequence and low-quality base calls, and aligned to the chromosome four reference sequence using Geneious software. **(a)** 5' RLM-RACE sequencing identified a single putative transcriptional start site for lncRNA-TARE-3L corresponding to the TARE 3 subtelomeric repeat boundary. Purple arrows represent the 5' RLM-RACE adapter/primer sequence (Ambion). **(b)** 3' RACE demonstrated that a long transcript encompassing the TARE2 subtelomeric repeat is transcribed towards the telomeres. Green arrows represent the 3' RACE adapter/primer sequence (Ambion). **(c)** 3' RACE also predicted a shorter transcript terminating just upstream of TARE 2. Green arrows represent the 3' RACE adapter/primer sequence (Ambion). Purple arrows represent the lncRNA-TARE family-specific 3' RACE primer sequence. Figure produced using Geneious v4.8 software.



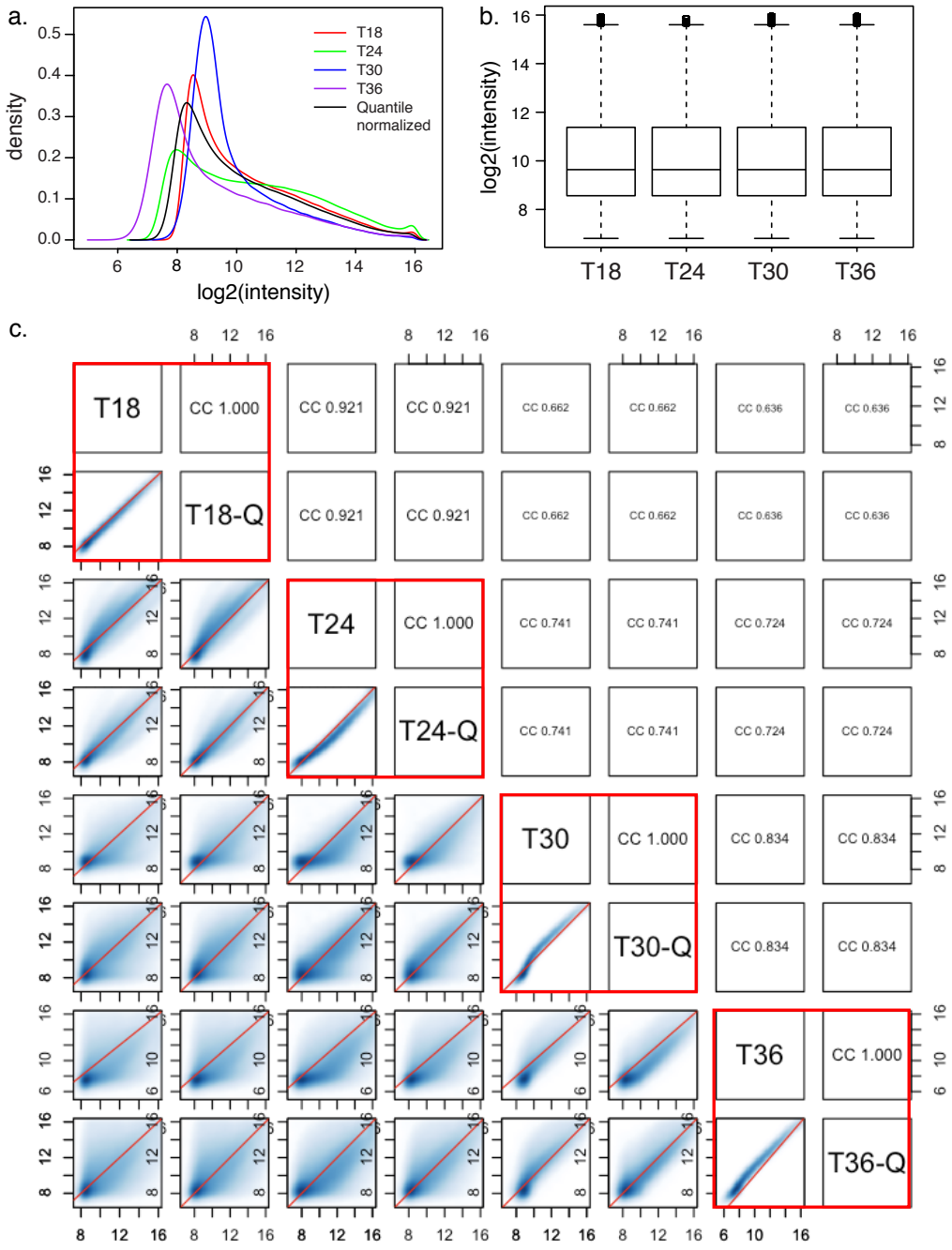
Supplementary Figure 2.5 (Continued) | Sequenced lncRNA-TARE RLM-RACE products.

Supplementary Figure 2.5 (Continued) | Sequenced lncRNA-TARE RLM-RACE products.

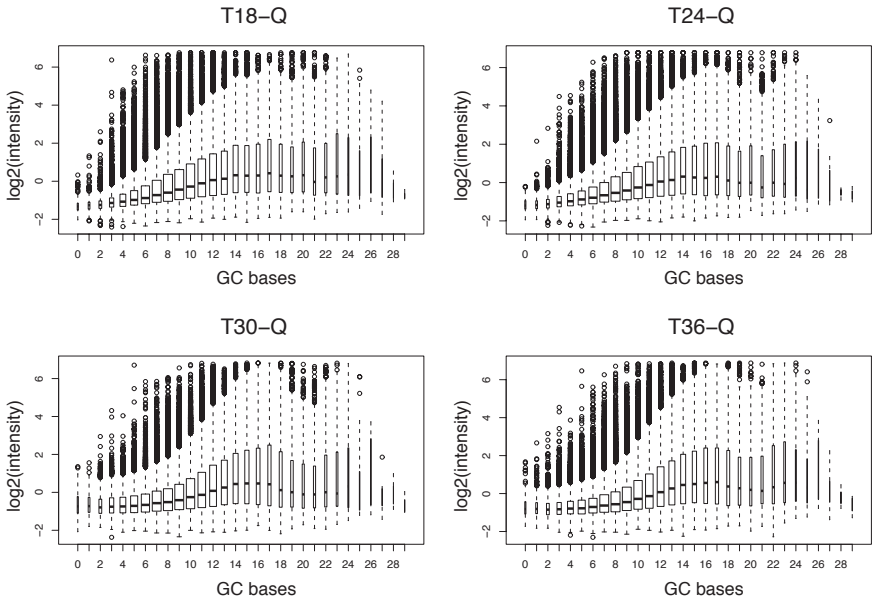
Supplementary Figure 2.5 (Continued) | Sequenced lncRNA-TARE RLM-RACE products



Supplementary Figure 2.6 | LncRNA-TARE is co-expressed with SPE2-binding transcription factor PfSip2. qRT-PCR amplification of lncRNA-TARE-4L and the SPE2-binding transcription factor PfSip2 demonstrates co-expression during parasite schizogony, with maximal lncRNA-TARE-4L expression occurring slightly after PfSip2 expression. LncRNA-TARE-4L and PfSip2 relative expression is plotted in time-points T30, T36, T38, and T40 +/- 3 hpi with respect to T30. Error bars represent the propagated standard error of the mean (pSEM) from four technical replicates. Housekeeping gene PF08_0085 was used as the reference gene.



Supplementary Figure 2.7 | Quantile normalization corrects systematic biases in raw data without overcorrecting. (a) Log2 probe hybridization intensity distributions from each sample prior to quantile normalization and after quantile normalization (black curve). **(b)** Boxplots of quantile normalized, log2 transformed probe hybridization intentensities from each sample. **(c)** Pairwise correlation scatterplots between raw and quantile normalized data (-Q) shows that the correlation is very strong between matched samples (Pearson's correlation coefficient (CC) = 1.0).



Supplementary Figure 2.8 | Minimal increase in median intergenic probe hybridiization intensity is observed with number of GC bases. Boxplots of intergenic probe hybridization intensities for each sample after quantile normalization. Probes are grouped according to the number of GC bases in their sequence.

Supplementary Table 2.1 | Stage-specific gene ontology analysis. The top four over-represented GO terms in each time-point.

	Best GOs	p-value	description
T18	GO:0020013	2.84E-08	rosetting
	GO:0007157	2.84E-08	heterophilic cell-cell adhesion
	GO:0020030	6.45E-08	infected host cell surface knob
	GO:0016337	6.45E-08	cell-cell adhesion
T24	GO:0044237	0.000131586	cellular metabolic process
	GO:0044238	0.000131586	primary metabolic process
	GO:0043170	0.003819134	macromolecule metabolic process
	GO:0006139	0.004118201	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
T30	GO:0016788	0.0321	hydrolase activity, acting on ester bonds
	GO:0004518	0.0624	nuclease activity
	GO:0006259	0.0755	DNA metabolic process
	GO:0006260	0.0933	DNA replication
T36	GO:0008234	0.112	cysteine-type peptidase activity
	GO:0008233	0.112	peptidase activity
	GO:0004197	0.112	cysteine-type endopeptidase activity
	GO:0004175	0.112	endopeptidase activity

Supplementary Table 2.2 | Putative small RNAs. Chromosome, coordinates, and length of 46 putative small RNAs eliminated from lncRNA candidate list.

chromosome	start	stop	length
Pf3D7_02	76857	76947	91
Pf3D7_02	91072	91184	113
Pf3D7_02	102873	102989	117
Pf3D7_02	175192	175358	167
Pf3D7_02	210863	210907	45
Pf3D7_02	273406	273534	129
Pf3D7_02	349389	349559	171
Pf3D7_02	410233	410375	143
Pf3D7_02	413194	413310	117
Pf3D7_02	441867	441911	45
Pf3D7_03	122490	122642	153
Pf3D7_04	269938	270032	95
Pf3D7_04	458795	458877	83
Pf3D7_04	627801	627849	49
Pf3D7_04	679162	679206	45
Pf3D7_04	704637	704775	139
Pf3D7_04	946452	946500	49
Pf3D7_04	958351	958399	49
Pf3D7_04	981177	981225	49
Pf3D7_05	1139348	1139540	193
Pf3D7_05	1140246	1140416	171
Pf3D7_07	208840	208912	73
Pf3D7_07	208992	209040	49
Pf3D7_07	807151	807233	83
Pf3D7_07	913847	914023	177
Pf3D7_07	995858	995964	107
Pf3D7_07	1068513	1068641	129
Pf3D7_09	133451	133495	45
Pf3D7_09	133589	133721	133
Pf3D7_09	363973	364045	73
Pf3D7_09	537936	538101	166
Pf3D7_09	650017	650201	185
Pf3D7_09	650371	650439	69
Pf3D7_09	700267	700371	105
Pf3D7_09	1042227	1042317	91
Pf3D7_09	1148725	1148817	93
Pf3D7_09	1254633	1254727	95
Pf3D7_09	1263209	1263337	129
Pf3D7_09	1331291	1331337	47
Pf3D7_09	1422839	1422883	45
Pf3D7_09	1427281	1427363	83
Pf3D7_09	1436275	1436333	59
Pf3D7_12	91081	91129	49
Pf3D7_12	163021	163219	199
Pf3D7_12	421028	421222	195
Pf3D7_12	577188	577306	119

Supplementary Table 2.3 | Putative genes or pseudogenes. Chromosome, coordinates, and length for 17 putative genes or pseudogenes eliminated from lncRNA candidate list.

chromosome	start	end	length
Pf3D7_02	10394	11762	1369
Pf3D7_02	86173	86529	357
Pf3D7_02	102270	102618	349
Pf3D7_04	224998	225306	309
Pf3D7_04	568293	568535	243
Pf3D7_07	80066	80462	397
Pf3D7_07	83609	84272	664
Pf3D7_07	774410	775089	680
Pf3D7_07	790877	791737	861
Pf3D7_07	1023527	1023988	462
Pf3D7_07	1336280	1336808	529
Pf3D7_09	177874	178908	1035
Pf3D7_09	537284	537678	395
Pf3D7_09	1138997	1139900	904
Pf3D7_09	1285576	1285999	424
Pf3D7_12	186108	186825	718
Pf3D7_12	195558	197020	1463

Supplementary Table 2.4 | Characteristics of 60 *P. falciparum* lncRNA candidates.
Properties of lncRNA-TARE-4L are boxed. Candidates highlighted in pink are located > 1 kb away from a neighboring protein-coding gene. Candidates highlighted in blue exhibit Pearson correlation < 0.9 with their more correlated neighboring gene.

	chr	start	stop	length	coverage (# of probes)	GC%	overlap with PlasmoDB v7.1 ncRNAs	nearest gene
1	Pf3D7_02	90364	90872	509	31	13.75%		PFB0095c
2	Pf3D7_02	98908	99480	573	30	9.77%		PFB0095c
3	Pf3D7_02	108948	109314	367	22	13.90%		PFB0105c
4	Pf3D7_02	161261	161889	629	19	11.29%		PFB0160w
5	Pf3D7_02	407481	408619	1139	60	15.89%	PF02TR001	PFB0445c
6	Pf3D7_02	410525	410755	231	13	22.94%		PFB0450w
7	Pf3D7_02	521868	522114	247	11	14.98%		PFB0560w
8	Pf3D7_02	593316	594090	775	48	15.74%		PFB0665w
9	Pf3D7_02	637779	638182	404	23	15.59%		PFB0700c
10	Pf3D7_02	638799	639127	329	16	13.98%		PFB0705w
11	Pf3D7_03	128996	129694	699	52	19.17%	PF03TR002	PFC0115c
12	Pf3D7_03	146103	146672	570	40	15.61%		PFC0130c
13	Pf3D7_04	8254	12006	3753	50	32.11%		PFD0005w
14	Pf3D7_04	156608	157411	804	51	20.90%		PFD0115c
15	Pf3D7_04	221362	222522	1161	51	12.75%	RNAzID:3065	PFD0185c
16	Pf3D7_04	465979	466243	265	20	20.75%		PFD0485w
17	Pf3D7_04	654078	654348	271	18	15.50%		PFD0692c
18	Pf3D7_04	707195	708303	1109	54	14.70%		PFD0775c
19	Pf3D7_04	735375	735781	407	20	14.99%	PF04TR001	PFD0807c
20	Pf3D7_04	830465	833119	2655	100	12.50%	PF04TR003	PFD0885c
21	Pf3D7_04	1055726	1055949	224	9	16.96%		PFD1100c
22	Pf3D7_04	1074520	1075252	733	37	17.74%		PFD1120c
23	Pf3D7_04	1119682	1120248	567	29	11.64%		PFD1170c
24	Pf3D7_05	1002403	1002926	524	13	9.54%		PFE1195w
25	Pf3D7_05	1018995	1019384	390	18	11.03%		PFE1230c
26	Pf3D7_05	1162996	1163395	400	13	13.25%		PFE1405c
27	Pf3D7_05	1170633	1171342	710	36	14.08%		PFE1415w
28	Pf3D7_05	1224183	1224589	407	29	16.46%		PFE1500c
29	Pf3D7_07	91404	92007	604	39	15.07%	PF07TR001	MAL7P1.230
30	Pf3D7_07	97857	98495	639	43	9.23%		MAL7P1.231
31	Pf3D7_07	105724	106082	359	27	13.37%		MAL7P1.228
32	Pf3D7_07	159485	159827	343	16	7.87%		PF07_0008
33	Pf3D7_07	271123	271401	279	11	16.49%		MAL7P1.19
34	Pf3D7_07	340595	340821	227	16	19.82%		MAL7P1.300
35	Pf3D7_07	370441	370752	312	10	16.67%		MAL7P1.204
36	Pf3D7_07	440174	440588	415	24	13.98%		PF07_0029
37	Pf3D7_07	1147335	1148031	697	48	18.79%		PF07_0107
38	Pf3D7_09	151334	151676	343	26	15.74%		PFI0165c
39	Pf3D7_09	275014	275478	465	27	13.76%		PFI0265c
40	Pf3D7_09	324073	325747	1675	55	9.85%	PF09TR001	PFI0325c
41	Pf3D7_09	326059	327734	1676	62	11.99%	PF09TR001	PFI0325c
42	Pf3D7_09	356053	356337	285	17	17.89%		PFI0360c
43	Pf3D7_09	440457	440863	407	24	17.69%		PFI0470w
44	Pf3D7_09	509630	509878	249	15	22.09%		PFI0545w
45	Pf3D7_09	554392	555099	708	31	15.11%		PFI0610w
46	Pf3D7_09	698206	699397	1192	44	12.84%		PFI0820c
47	Pf3D7_09	741852	742164	313	22	18.53%		PFI0880c
48	Pf3D7_09	762853	763259	407	25	24.57%	RNAzID:4653	PFI0905w
49	Pf3D7_09	852079	852445	367	16	12.53%		PFI1020c
50	Pf3D7_09	939059	939539	481	27	17.67%		PFI1130c
51	Pf3D7_09	1074703	1075688	986	55	16.84%	RNAzID:4694	PFI1300c
52	Pf3D7_09	1396100	1396574	475	20	16.21%		PFI1715w
53	Pf3D7_09	1426449	1426655	207	13	15.46%		PFI1740c
54	Pf3D7_09	1426665	1427132	468	20	12.39%		PFI1740c
55	Pf3D7_09	1431827	1432370	544	29	18.57%	RNAzID:4797	PFI1745c
56	Pf3D7_12	235695	236219	525	23	14.67%		PFL0270c
57	Pf3D7_12	304588	305687	1100	64	13.45%		PFL0335c
58	Pf3D7_12	413796	415379	1584	59	12.63%		PFL0445w
59	Pf3D7_12	594179	594923	745	38	13.15%		PFL0675c
60	Pf3D7_12	598926	599486	561	32	14.08%		PFL0675c

Supplementary Table 2.4 (Continued) | Characteristics of 60 *P. falciparum* lncRNA candidates.

	distance to nearest gene	max correlation with neighboring gene	fold change	distance to neighboring gene with max correlation	max time-point	max probe intensity
1	446	0.995052594	3.67727881	446	T18	25.3
2	70	0.981759536	3.342655313	70	T18	6.1
3	250	0.97958822	1.841013208	250	T36	6.5
4	19	0.938022972	1.571523106	176	T24	7.4
5	1126	0.759336454	1.661183541	1126	T18	3.0
6	297	-0.167377667	0.273653001	297	T24	37.0
7	122	0.099466422	2.114153871	122	T24	17.8
8	106	0.995196901	3.169826237	106	T36	3.2
9	272	0.996325865	0.822834061	272	T18	5.2
10	1273	0.993359742	0.784461262	1292	T24	9.1
11	2081	0.964970012	2.408835642	2403	T36	2.7
12	119	0.974126719	1.677265219	119	T18	4.1
13	23147	-0.295570023	2.591987191	39996	T36	2.8
14	1768	0.783888023	2.231480862	1768	T18	3.1
15	1044	0.976884288	1.963195203	1044	T36	3.4
16	364	0.619950843	1.003855895	364	T18	6.3
17	24	0.999387225	3.732924308	24	T30	9.6
18	206	-0.151776699	0.273846694	206	T36	4.5
19	413	0.997818747	0.757552646	413	T24	6.0
20	1144	0.952507595	0.904543821	1144	T24	3.5
21	67	0.95271842	2.55613031	67	T36	5.3
22	621	0.828742942	1.731965884	621	T24	12.0
23	1567	0.962824823	2.906810269	3058	T18	5.2
24	284	0.669843589	1.189718072	284	T24	10.2
25	196	0.878751018	1.494455992	196	T24	10.9
26	7	0.839679446	2.27533969	1449	T24	15.3
27	311	0.980654158	3.103472129	3086	T36	3.8
28	2	0.626834415	1.999431118	2	T24	9.6
29	128	0.994484972	2.753246062	128	T18	9.3
30	176	0.917669192	2.801368561	4888	T18	3.4
31	142	0.872494021	2.213895188	142	T24	6.2
32	496	0.989753519	3.471594294	496	T18	19.9
33	3	0.999824531	2.152716093	3	T18	14.8
34	32	0.947310662	3.352449172	32	T18	57.1
35	40	0.993695202	2.28866887	40	T18	11.0
36	182	0.977057716	1.835105868	373	T18	7.4
37	911	0.725560466	2.104869294	911	T18	3.3
38	23	0.991803058	2.012729887	23	T18	18.2
39	227	0.953898382	2.667464921	227	T36	5.7
40	734	0.990173149	2.215246438	734	T18	3.1
41	2720	0.986217308	1.299729826	2720	T18	2.5
42	321	0.938508778	2.904288987	321	T36	11.6
43	43	-0.487368854	1.222003636	5448	T18	5.5
44	47	0.967111263	2.042565141	47	T36	10.4
45	982	0.950320895	1.916811529	982	T18	4.0
46	187	0.965692296	1.110192479	187	T18	2.9
47	320	0.915651649	3.44931118	320	T36	9.1
48	137	0.604987619	1.924518765	137	T24	6.4
49	70	0.945066028	2.687622707	70	T18	15.1
50	31	0.968663876	1.402635855	31	T24	9.4
51	217	0.991764486	1.50043645	876	T36	2.3
52	4215	0.110732928	0.809282807	16359	T30	7.3
53	808	0.998617666	3.474515586	808	T18	11.3
54	331	0.979374432	4.012441607	331	T18	20.7
55	692	0.775882784	1.962751937	692	T18	4.3
56	51	0.304466604	0.173562662	51	T24	5.1
57	1765	0.976556071	2.024865895	1765	T18	3.4
58	19	-0.422811312	0.459209104	3322	T24	2.7
59	237	0.967954798	0.805568773	237	T18	11.5
60	289	0.972201999	1.440896368	289	T18	3.6

Supplementary Table 2.5 | qRT-PCR and RLM-RACE primers. Gene-specific lncRNA-TARE/PfSip2 (PFF0200c) qRT-PCR primer sequences and family-specific lncRNA-TARE RLM-RACE primer sequences.

target	sequence
lncRNA-TARE-1L F	ATAACCGTACTACTTCCT
lncRNA-TARE-1L R	CATCATATTATGTTGAAGCTA
lncRNA-TARE-2R F	ACTTACGAACTAAGGAAA
lncRNA-TARE-2R R	AATGATGATGATTTGGG
lncRNA-TARE-3L F	TACTACTTCCTAGCAACA
lncRNA-TARE-3L R	CATGAAATAGAATATGGGTTAA
lncRNA-TARE-3R F	TTTGTTTGCATTATGTGTT
lncRNA-TARE-3R R	AACCATTTAGTCCAATCAA
lncRNA-TARE-4L F	AATTTACCCATTTTTCAAA
lncRNA-TARE-4L R	AAGTACTCGTGGCAAT
lncRNA-TARE-4R F	ACTTCTAACTATCAGATTCAT
lncRNA-TARE-4R R	AACCATATCATCATTAAGAAC
lncRNA-TARE-6R F	TTTCCTAAAAAATTTGGTCCAT
lncRNA-TARE-6R R	CGTGGCAATTAGGTGAA
lncRNA-TARE-7L F	GGGTAGTACCATAAGTTT
lncRNA-TARE-7L R	TACAACTAACCTAACTAAC
lncRNA-TARE-7R F	AACTGGAGGAATAAATTCT
lncRNA-TARE-7R R	GTTGTTCGGAAGTAGTAA
lncRNA-TARE-8R F	TCTAACTATCACATTCATATCT
lncRNA-TARE-8R R	CATCTACAACATCATCAATG
lncRNA-TARE-9L F	GTGTGTTTGTTTGGGGGTAG
lncRNA-TARE-9L R	AACTAAGCCAACTAACAACATCAA
lncRNA-TARE-9R F	ATGGTCACTTACAAACTAAT
lncRNA-TARE-9R R	GTGTGTTTTGGGGTTAA
lncRNA-TARE-10L F	TTCTAATACTACTACTTCATGT
lncRNA-TARE-10L R	GTGGTAAATAATGGAGGAA
lncRNA-TARE-11L F	TGTGAAGGGGTATTAC
lncRNA-TARE-11L R	AATTGGTGAACTATGAC
lncRNA-TARE-12L F	TTAGTTCTACCCGTTTG
lncRNA-TARE-12L R	CACATCAAGCAGACAT
PfSip2 F	GTACCATTAAATTCACGTAAGG
PfSip2 R	TATTATTATTCTCCATACCATCCC
5'-RACE-outer	CACTAAGGTTCAAGGTTCTC
5'-RACE-inner	GATGTTGTAGATGATGTCGTT
3'-RACE	GAGAACCTTGAACCTTAGTG

Supplementary Table 2.6 | RLM-RACE syntenic transcript coordinates. Original chromosome and syntenic coordinates on the left end of chromosome four for each sequenced RACE product.

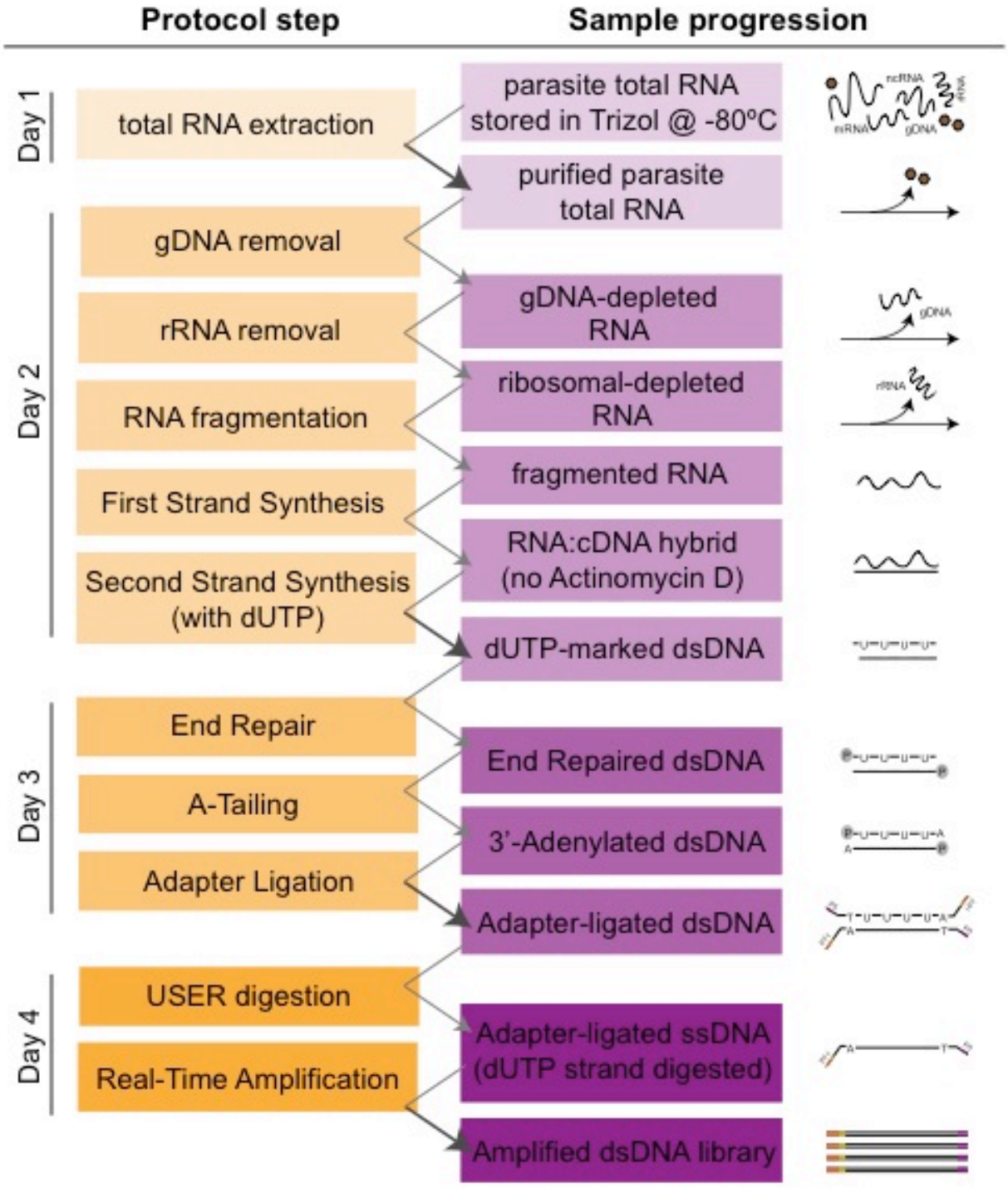
3' RACE products

original chromosome (any ambiguities are noted)	syntenic coordinates on chromosome four
2 right	8156
5 left or 13 left	8156
4 left or 6 right	8209
9 right	8210
4 left or 6 right	8213
5 left or 13 left	8221
8 left	8221
12 left	8221
2 right	8221
9 right	8221
1 left	8221
1 left	8221
4 left or 6 right	8221
9 right	8221
1 left	9821
4 left or 6 right	9821
5 left or 13 left	9821
3 right	9821
7 right	9821
7 right	9821
7 left	9821
2 right	9821
9 right	9826
4 left or 6 right	9875
3 right	9958
3 right	10022
12 left	10050

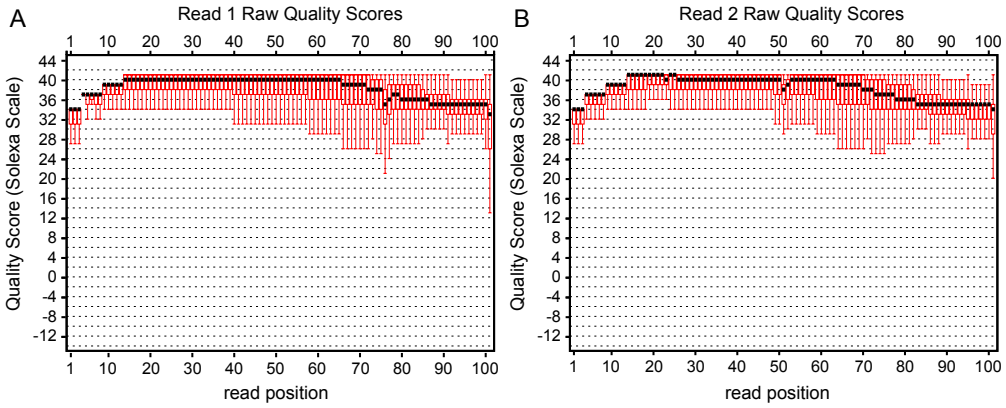
5' RACE products

original chromosome (any ambiguities are noted)	syntenic coordinates on chromosome four
3 left	11278
3 left	11278
3 left	11278
3 left	11278
3 left	11278
3 left	11278
3 left	11278
3 left	11278
3 left	11278
3 left	11278

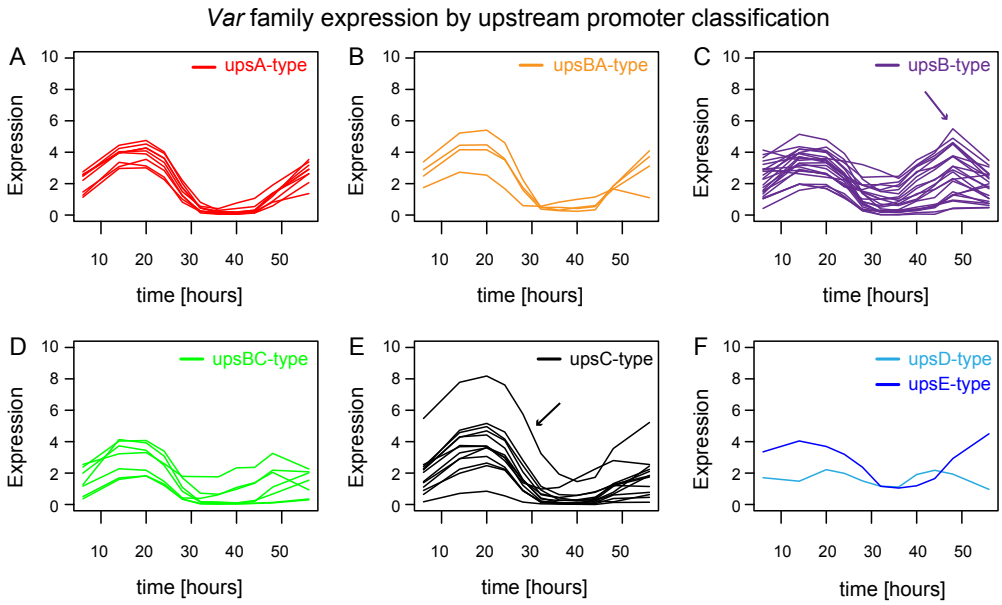
Chapter 3 Supplementary Figures, Tables, and Protocol



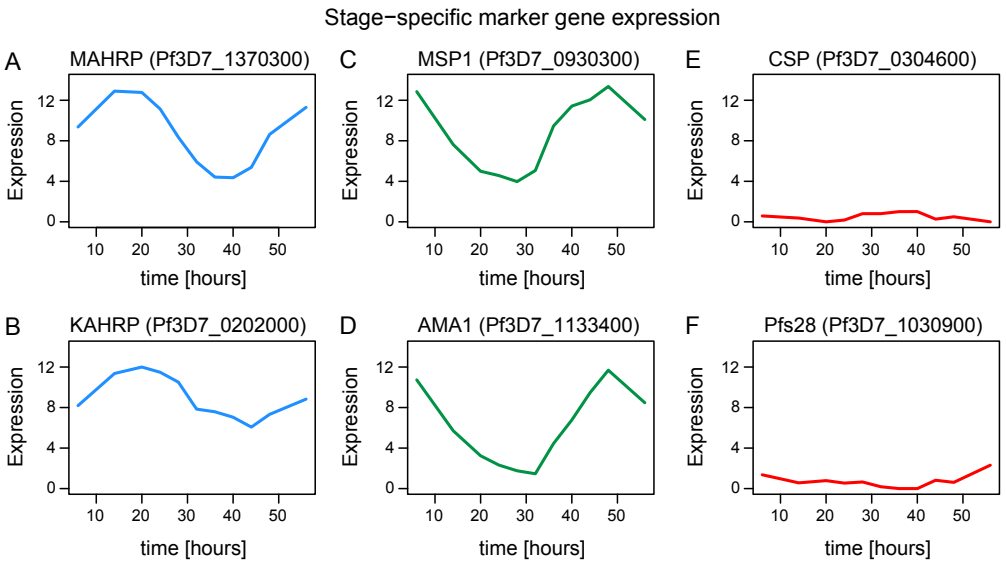
Supplementary Figure 3.1 | Strand-specific, non-polyA-selected *P. falciparum* libraries can be comfortably prepared in one week. The flowchart shows major protocol steps and sample progression during library preparation, beginning with parasite total RNA stored in TRIZOL Reagent and ending with amplified double-stranded DNA libraries ready for quantification and Illumina sequencing.



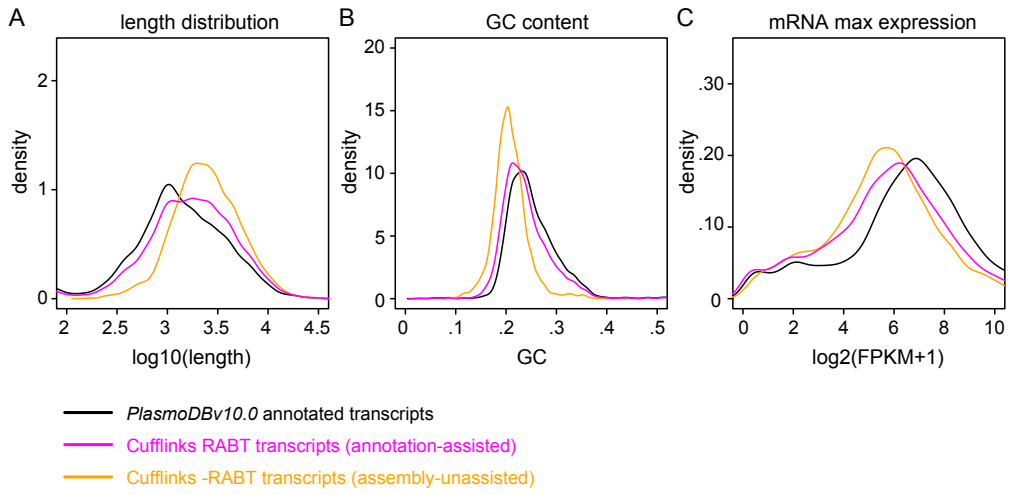
Supplementary Figure 3.2 | Raw reads have high quality scores. (A/B) Read 1 and Read 2 raw quality scores, respectively. Each plot summarizes approximately 307 million reads agglomerated from all fifteen samples. We trimmed the last base and filtered the 4-8% of reads from each sample that did not pass Illumina filtering prior to read alignment.



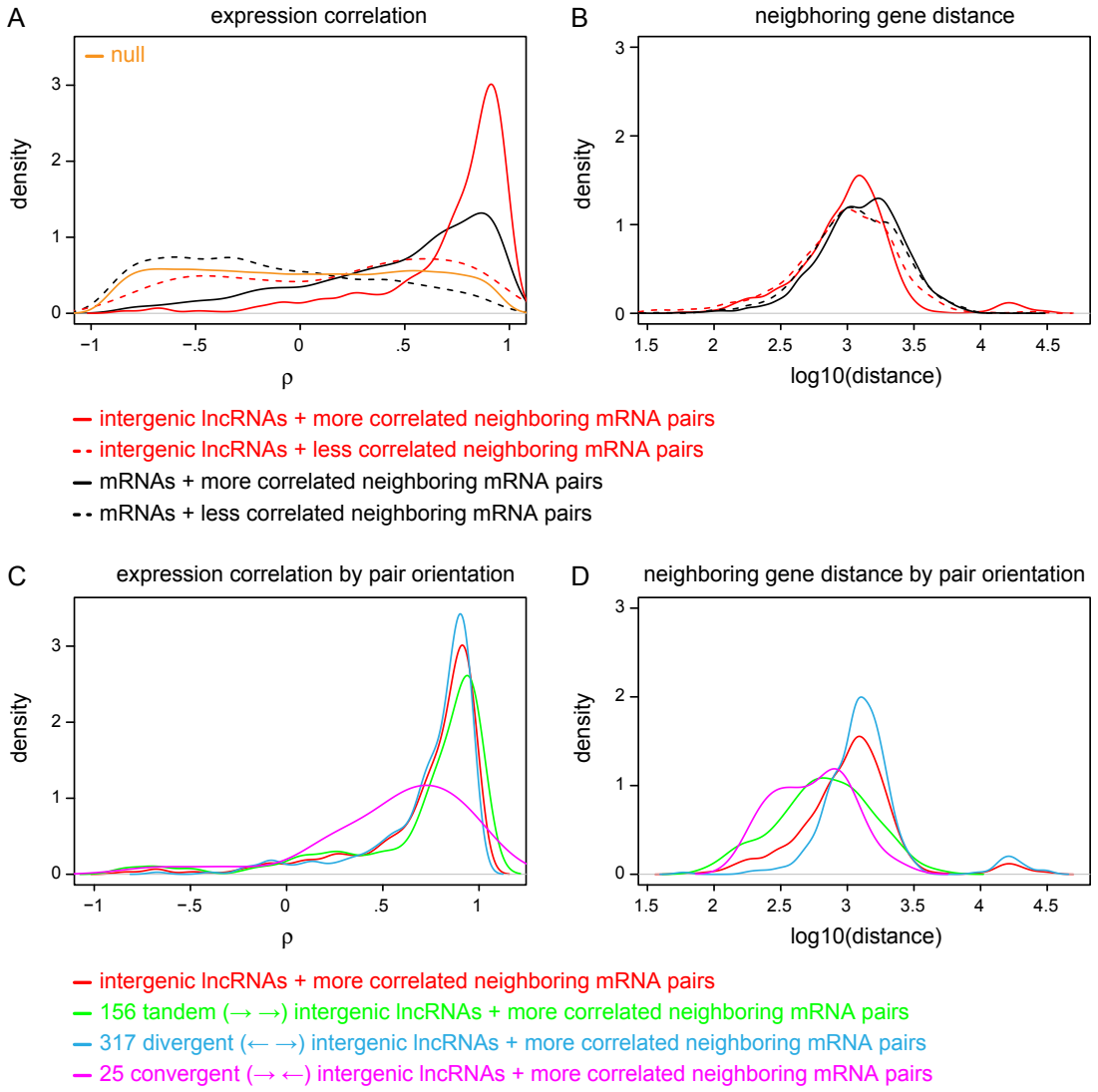
Supplementary Figure 3.3 | A central upsC-type var gene is the dominant var transcript, and subtelomeric upsB-type var genes produce the most non-coding transcripts. **(A/B/C/D/E/F)** Expression during the 56-hour time course of upsA-type (red), upsBA-type (orange), upsB-type (purple), upsBC-type (green), upsC-type (black), upsD-type (light blue) and upsE-type (dark blue) var genes, respectively [1]. This analysis indicated that an upsC-type gene, Pf3D7_0412700, was the dominant var transcript in the parasite population (black arrow). We also noted that the transcription of non-coding transcripts varied by upstream promoter (ups) type. Interestingly, the majority of upsB-type var genes transcribed non-coding transcripts during parasite invasion (purple arrow), whereas the majority of upsA-type and upsC-type var genes did not. Expression is plotted in units of log2(FPKM+1). upsA-type: Pf3D7_0425800, Pf3D7_1150400, Pf3D7_1300300, Pf3D7_0800200, Pf3D7_1100200, Pf3D7_0400400, Pf3D7_0100300, Pf3D7_0937600, Pf3D7_0600400; upsBA-type: Pf3D7_0800300, Pf3D7_1200400, Pf3D7_0600200, Pf3D7_0632500; upsB-type: Pf3D7_1100100, Pf3D7_0800100, Pf3D7_0500100, Pf3D7_0100100, Pf3D7_0115700, Pf3D7_0324900, Pf3D7_0400100, Pf3D7_0900100, Pf3D7_0223500, Pf3D7_1041300, Pf3D7_1200100, Pf3D7_0200100, Pf3D7_0300100, Pf3D7_1255200, Pf3D7_0426000, Pf3D7_0937800, Pf3D7_1000100, Pf3D7_1219300, Pf3D7_0632800, Pf3D7_1373500, Pf3D7_0733000, Pf3D7_0700100, Pf3D7_1300100, Pf3D7_0833500; upsBC-type: Pf3D7_0413100, Pf3D7_1240400, Pf3D7_0808700, Pf3D7_0712800, Pf3D7_0712400, Pf3D7_0421100, Pf3D7_1240300; upsC-type: Pf3D7_0712900, Pf3D7_0808600, Pf3D7_0420700, Pf3D7_0712000, Pf3D7_0412900, Pf3D7_0420900, Pf3D7_0421300, Pf3D7_0412400, Pf3D7_0712600, Pf3D7_0711700, Pf3D7_1240600, Pf3D7_0412700, Pf3D7_0617400; upsD-type: Pf3D7_0533100; upsE-type: Pf3D7_1200600.



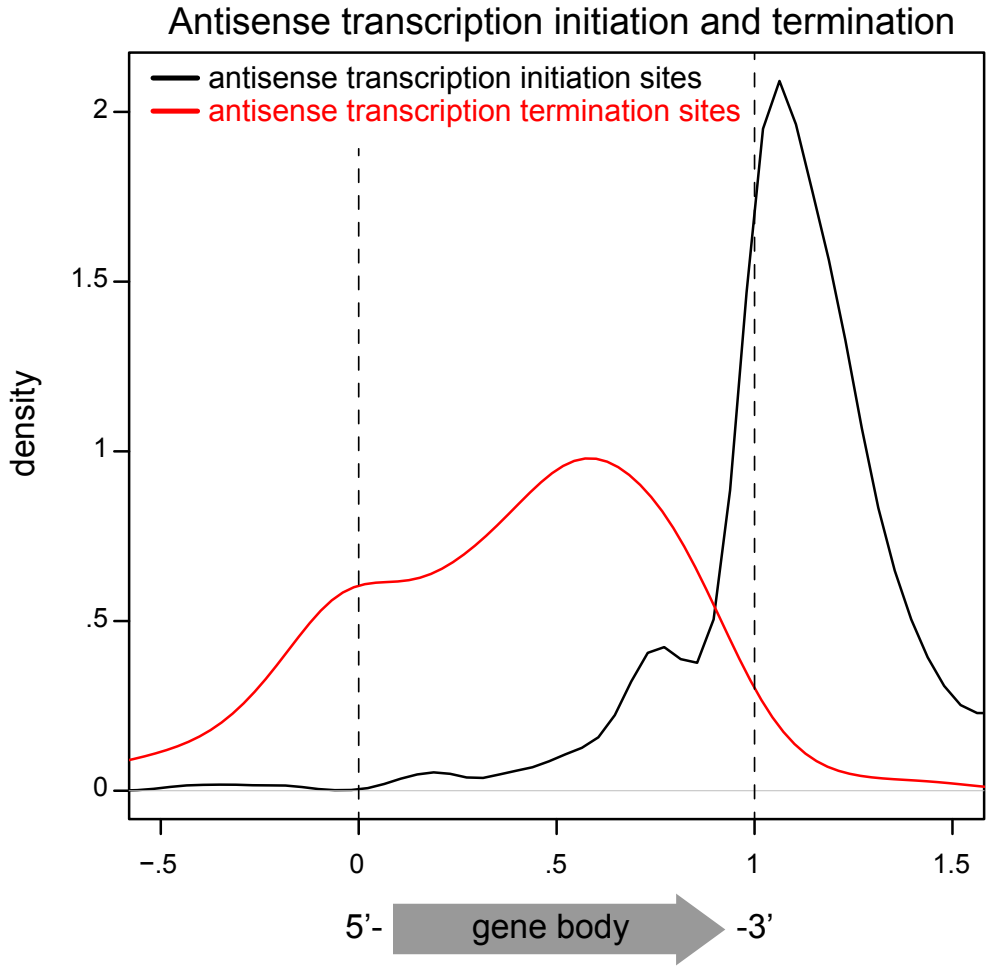
Supplementary Figure 3.4 | Stage-specific marker gene expression profiles match expectations. (A/B) Expression profiles of well established ring stage markers during the 56-hour time course. **(C/D)** Expression profiles of well established invasion markers during the 56-hour time course. **(E/F)** Expression profiles of well-established mosquito stage markers during the 56-hour time course. MAHRP = membrane-associated histidine rich protein; KAHRP = knob-associated histidine rich protein; MSP1 = merozoite surface protein 1; AMA1 = apical membrane antigen 1; CSP = circumsporozoite protein; Pfs28 = 28 kDa ookinete surface protein.



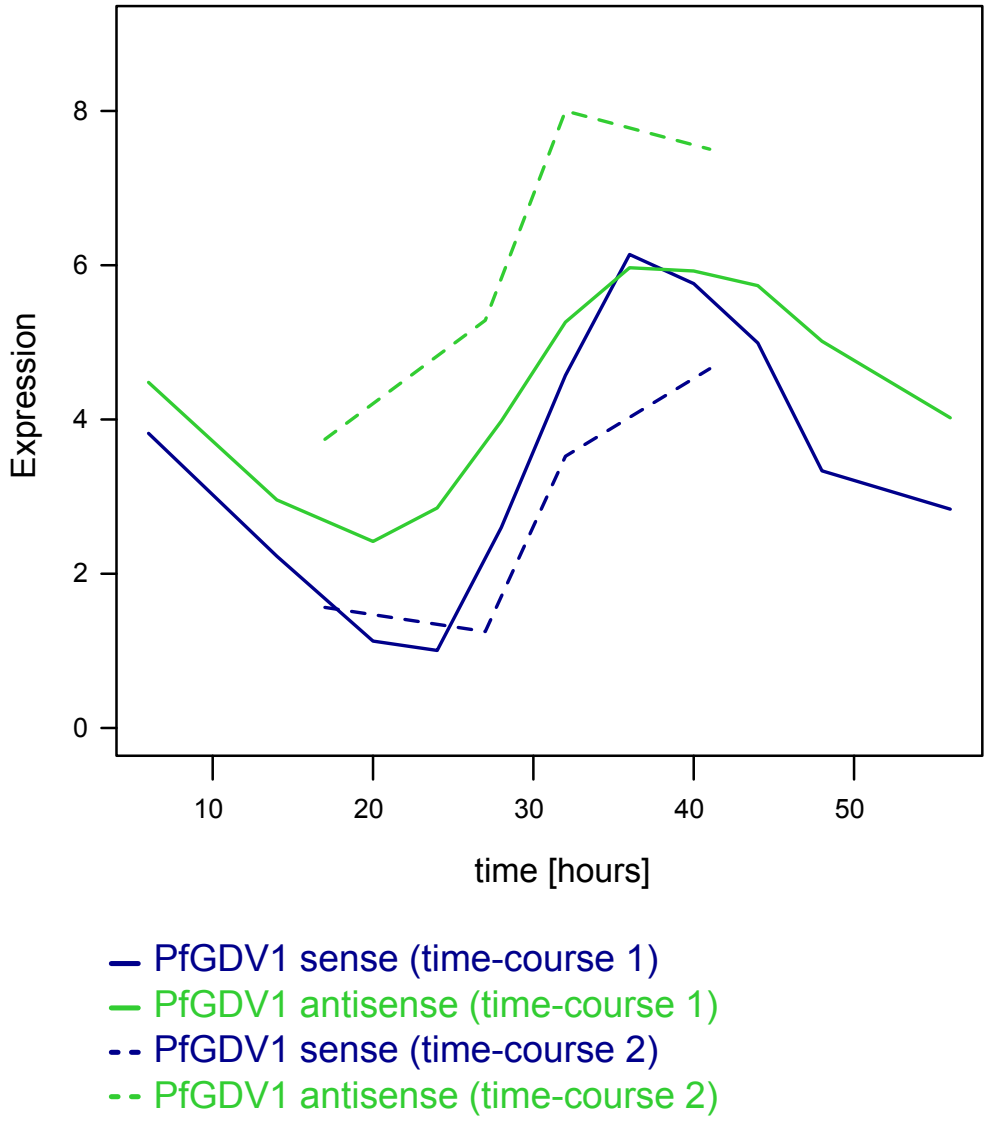
Supplementary Figure 3.5 | Cufflinks assembled transcripts tend to include un-annotated regions, likely to be UTRs. Upon comparing the properties of PlasmoDBv10.0 annotated transcripts (black), Cufflinks RABT transcripts (pink), and Cufflinks -RABT transcripts (orange), we found evidence for un-annotated, potential UTRs included in the Cufflinks assembled transcripts. This was supported by the slightly (A) increased length and (B) reduced GC content of Cufflinks assembled transcripts, as P. falciparum non-coding regions typically have lower GC content than coding regions. Also, when we considered mRNA transcript expression levels during the 56-hour time course, we found (C) reduced maximum expression levels of Cufflinks assembled mRNA transcripts, which further indicated that Cufflinks transcripts included UTRs with less read coverage.



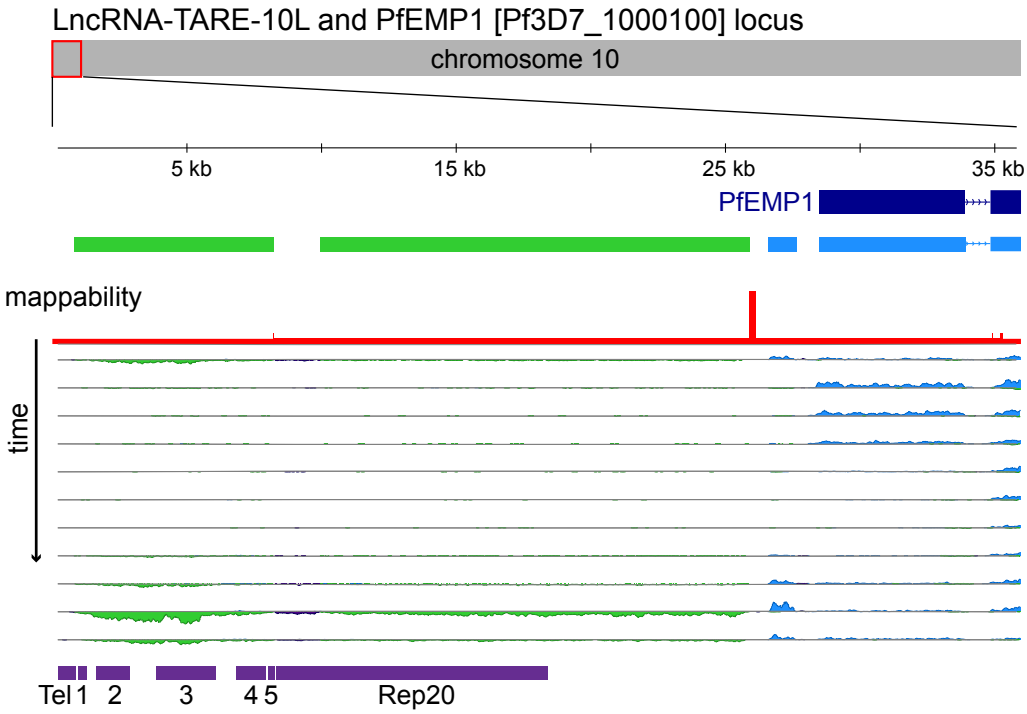
Supplementary Figure 3.6 | Two-dimensional genomic distance between neighboring loci is not a strong indicator of expression correlation. (A)/(B) The Pearson correlation during the 56-hour time course and the distance between mRNA-neighbor pairs (black) and intergenic lncRNA-neighbor pairs (red), respectively. Dashed lines correspond to pairings with the less correlated neighbor. This analysis indicated that both mRNAs and intergenic lncRNAs had a significantly more correlated neighbor. However, regardless of pairing criteria, the genomic distances between neighboring loci were quite similar. **(C/D)** We next stratified the Pearson correlation during the 56-hour time-course and the distance between intergenic lncRNAs and their more correlated neighbor by pair orientation, respectively. We found 156 tandem, 317 divergent, and 25 convergent pairs. Notably, convergent loci, albeit exhibiting less extreme positive correlation, were not more distant from each other.



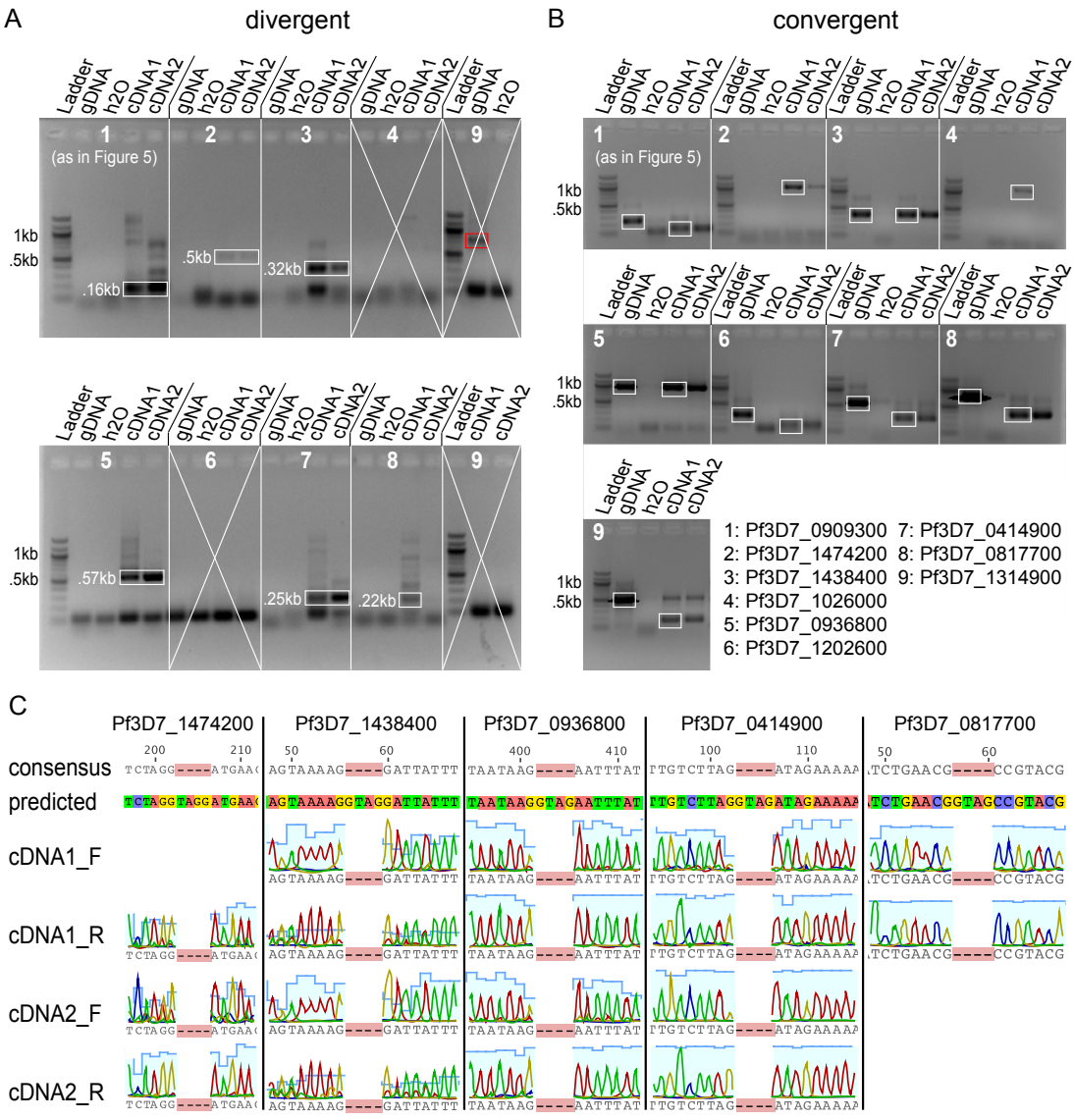
Supplementary Figure 3.7 | Sense-antisense transcripts tend to overlap in their 3' tail region. The distribution of antisense transcription initiation sites (black) and antisense transcription termination sites (red) showed that the vast majority of antisense transcripts initiated transcription downstream of gene bodies, and tended to terminate transcription towards the 3' tail region as well. We normalized antisense transcription initiation and termination sites to the location and length of their respective PlasmoDBv10.0 annotated gene partner, inclusive of intronic sequence.



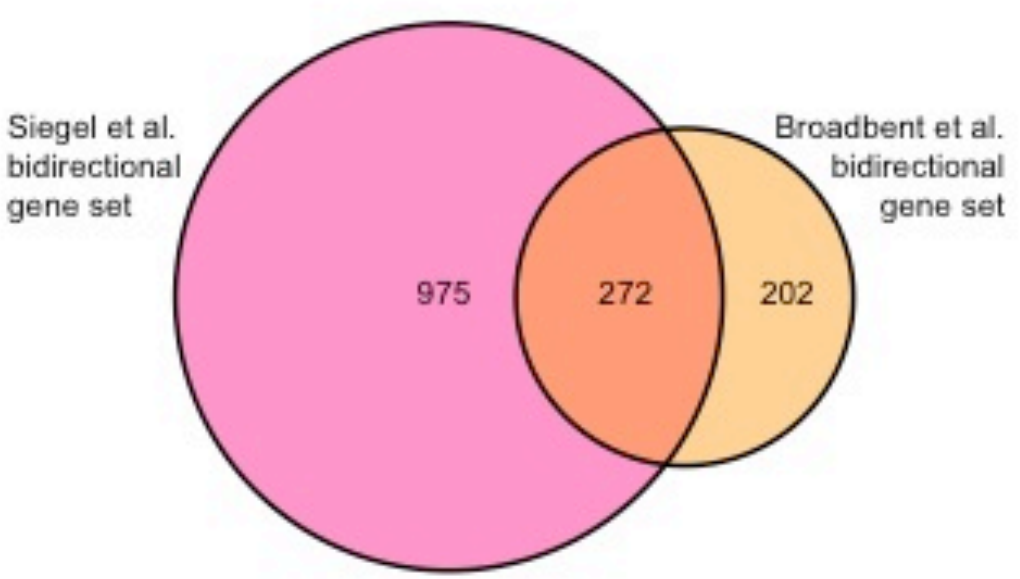
Supplementary Figure 3.8 | Comparison of PfGDV1 antisense transcript levels in two time courses. Plotting PfGDV1 expression during both time-courses showed that PfGDV1 sense-antisense transcript levels were highly correlated. Antisense transcript levels during the shorter time course reached a maximum FPKM of 255. Pearson correlation during the shorter time course was 0.85. We estimated the shorter time course time-points from our MDS staging analysis as approximately: 17, 27, 32, and 41 hpi. Expression is plotted in units of log2(FPKM+1).



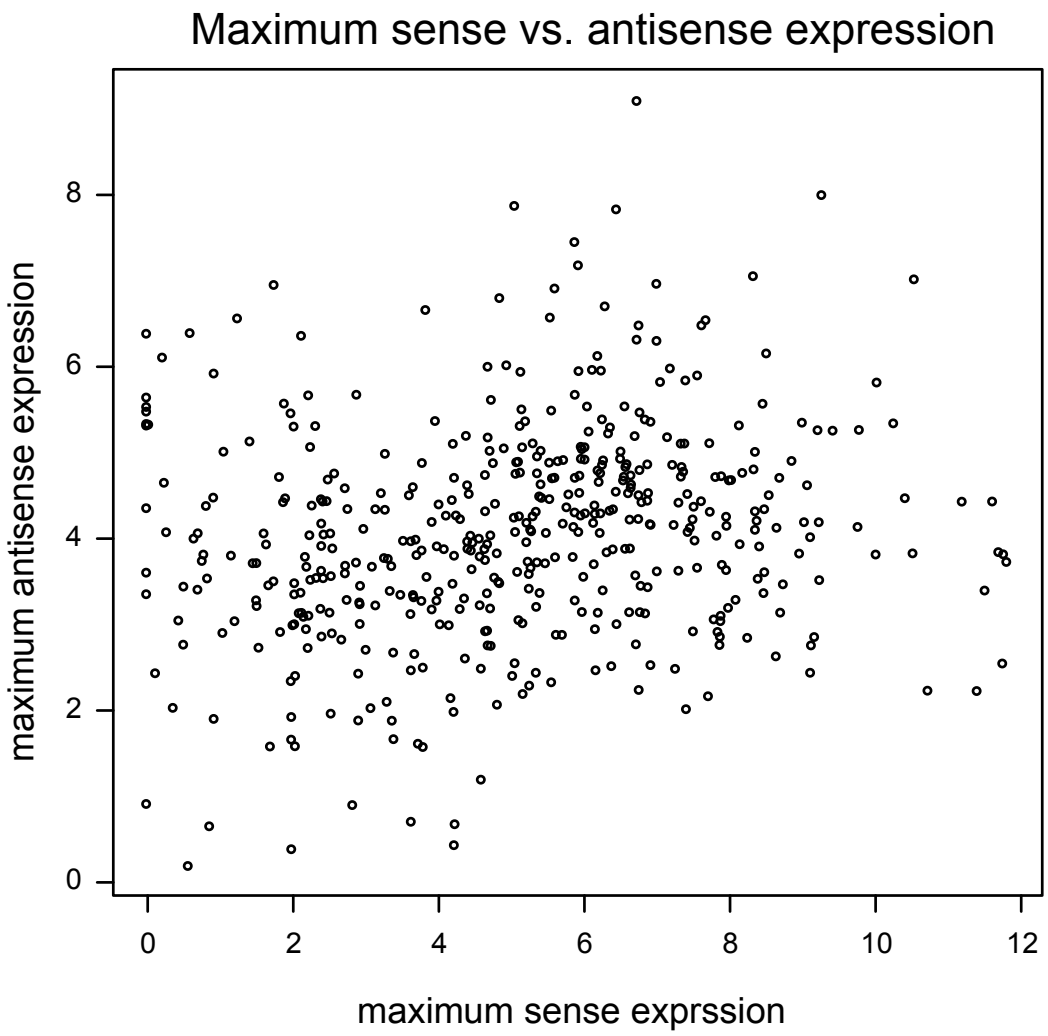
Supplementary Figure 3.9 | Transcription of the entire chromosome ten left subtelomere during invasion. Normalized read alignment tracks across the subtelomeric upsB-type PfEMP1 encoding var gene (Pf3D7_1000100) and chromosome ten left subtelomeric TARE region. The annotated Pf3D7_1000100 gene model is shown in dark blue, and assembled transcript models are shown in light green and light blue. Reads from each 56-hour time course sample mapping to the (-) strand are shown below each horizontal axis in light green, while reads mapping to the (+) strand are shown above each horizontal axis in light blue. Intron reads are shown in purple. Uniqueness of 100mers is plotted in red as a mappability track. Boundaries of the telomere, TAREs 1-5, and Rep20 are shown in purple as well.



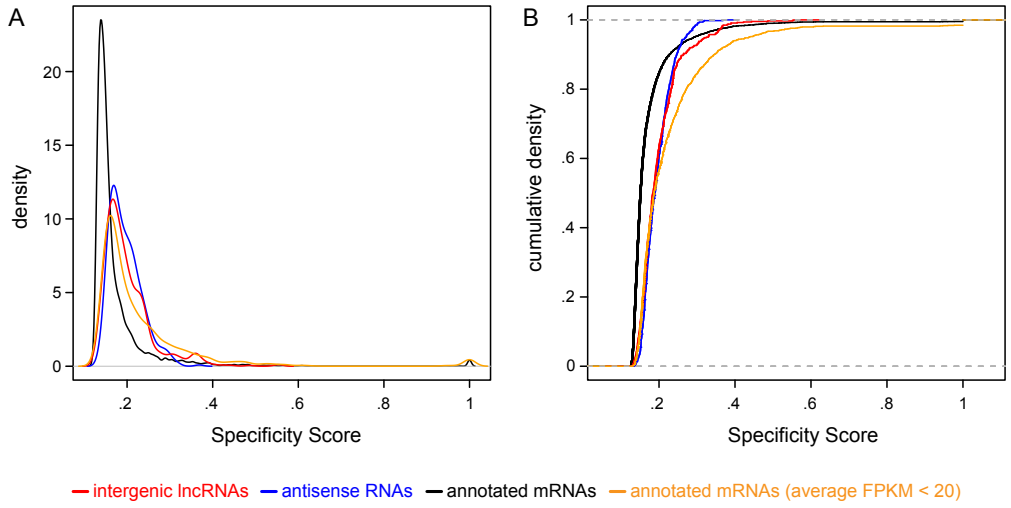
Supplementary Figure 3.10 | Experimental confirmation of five additional P. falciparum circRNAs. (A)/(B) Divergent and convergent PCR reactions, respectively, for the nine tested loci using gDNA, water (h20), and cDNA prepared from either time-course (cDNA1/2). Ladder and divergent amplicon sizes are indicated. We gel-excised and sequenced each boxed amplicon. White boxes indicate amplicons of the expected size with positive Sanger sequencing confirmation. The red box indicates a non-specific amplicon (according to sequencing results). Smaller amplicon sizes in the convergent cDNA reactions as compared to the convergent gDNA reactions corresponded to intron removal. The larger products in the divergent cDNA reactions may represent non-specific or rolling-circle reverse transcription products. **(C)** Sanger sequencing across five additional circRNA splice junctions.



Supplementary Figure 3.11 | Overlap with a previously described bidirectional gene set. Comparing our bidirectional (sense and antisense producing) gene set to the Siegel et al. bidirectional gene set indicated significant differences [2]. However, the Siegel et al. study only included four time-points, which could account for the 202 novel genes found here. The Siegel et al. study also did not involve assembly of transcript structures, and we required assembled transcript structures to be supported by at least fifty (perfectly and uniquely) mapped reads in a single sample. Our use of different methods and a higher threshold for expression may account for the 975 genes predicted by Siegel et al. only. Indeed, we calculated that the average maximum RPKM of "stringent" antisense regions reported by Siegel et al. was 4, while the average maximum FPKM of our antisense transcripts was 25. The FPKM (fragment) measurements reported by Cufflinks are analgous to single-end RPKM (read) measurements [3].



Supplementary Figure 3.12 | Maximum antisense expression level does not strongly correlate with maximum sense expression level. Each point represents a bidirectionally transcribed (sense and antisense) gene. We compared maximum sense and antisense transcript levels during the 56-hour time course. Expression is plotted in units of log2(FPKM+1).



Supplementary Figure 3.13 | Stage-specificity of P. falciparum lncRNAs as compared to mRNAs. (A) The distribution of specificity scores for each transcript class indicated that intergenic lncRNAs (red) and antisense RNAs (blue) typically had higher specificity scores than annotated mRNAs (black). However, when we considered only lowly expressed mRNAs, we found that the 1619 annotated mRNAs with an average FPKM of less than 20 (orange) had a comparable distribution of specificity scores to lncRNAs. **(B)** As in panel A, except that the cumulative distribution of specificity scores is plotted.

Supplementary Protocol 3.1 | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

0. DNASE TREATMENT

Reagents:

1. Up to 42 uL total RNA
2. Ambion TURBO DNase (order new lot, or ensure < 6 months old)
3. 10X TURBO DNase Buffer
4. Ambion SUPERase-In (20 U/uL)
5. Invitrogen RNaseOUT (40 U/uL)
6. Agencourt RNAclean SPRI beads
7. *Freshly prepared* 75% Ethanol
8. (optional) Agilent Bioanalyzer RNA 6000 Pico Kit

Protocol:

Decontaminate work space, work quickly, and maintain RNase-free

1. Combine the following in a PCR tube-strip:

___ uL	total RNA
___ uL	H_20
5 uL	10X TURBO DNase Buffer
1 uL	SUPERase-In
1 uL	RNaseOUT
1 uL	TURBO DNase
50 uL	

2. Incubate at room temperature for 30 minutes and proceed immediately to clean up to stop reaction
3. **1.8X RNAclean** SPRI bead purification in 96-well plate
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    b. Transfer 50 uL reaction to 96-well plate, add **90 uL beads**, and immediately pipette mixture 10 times upon bead addition
    c. Seal plate and incubate at room temperature for 20 minutes
    d. Separate beads on magnet for 5 minutes or until solution appears clear
    e. Remove and discard supernatant
    f. Keeping plate on magnetic rack, wash the beads in 200 uL freshly prepared 75% Ethanol for at least 30 seconds
    g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
    h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
    i. Resuspend beads in **30 uL h20** and incubate at room temperature for 5 minutes
    j. Separate beads on magnet for 5 minutes or until solution appears clear
    k. Carefully transfer 26-28 uL fragmented RNA to new tube/plate
4. Proceed immediately to Ribosomal RNA depletion
5. *Recommended:* save 1 uL aliquot of RNA to check RNA quality using an Agilent Bioanalyzer RNA 6000 Pico Kit

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

2. RIBOSOMAL RNA DEPLETION

Reagents:

1. 1-5 ug DNase-treated RNA (cleaned up, in RNase-free h20)
    a. 26 uL maximum volume (for 2.5-5 ug) or 28 uL for < 2.5 ug input
2. Epicentre Ribo-Zero Magnetic Kit (Human/Mouse/Rat)
3. Heating block for 1.5 mL tubes, **set to 50^{0}C**
4. Agencourt RNAclean SPRI beads
5. *Freshly prepared* 75% Ethanol
6. (optional) Agilent Bioanalyzer RNA 6000 Pico Kit

Protocol:

Decontaminate work space, work quickly, and maintain RNase-free

1. Equilibrate Ribo-Zero Magnetic Core Kit components to room temperature for 30 minutes
2. Remove the Ribo-Zero rRNA Removal Kit from -80^{0}C, thaw the tubes, and place on ice
3. Individual Magnetic Bead washing procedure:
    a. Mix the Ribo-Zero Magnetic Beads by gentle vortexing
    b. For each reaction, *slowly* pipette 225 uL of Ribo-Zero Magnetic Beads into a 1.5 mL RNase-free tube. Store unused beads at 4^{0}C
    c. Separate beads on magnet for 5 minutes or until solution appears clear
    d. With the tube still on the stand, remove and discard the supernatant
    e. Remove the tube from stand and add 225 uL of RNase-free h20. Mix well by vortexing at medium speed
    f. Repeat steps *c* and *d* and remove the tube from the magnetic stand
    g. Add 65 uL of Ribo-Zero Magnetic Bead Suspension Solution to each tube. Mix well by vortexing at medium speed
    h. Add 1 uL of RiboGuard RNase Inhibitor to each tube of resuspended Magnetic Beads. Mix briefly by vortexing.
    i. Store at room temperature until required in step *5*
4. Treatment of total RNA with Ribo-Zero rRNA Removal Solution
    a. In a PCR tube-strip combine the following:

_____ uL	h20
4 uL	Ribo-Zero Reaction Buffer
_____ uL	2.5-5 ug total RNA sample*
10 uL	Ribo-Zero rRNA Removal Solution
40 uL	

*If < 2.5 ug total RNA, adjust Ribo-Zero rRNA Removal Solution to 8 uL

    b. Gently mix the reaction(s) by pipetting and incubate at 68^{0}C for 10 minutes
    c. Store the remaining Ribo-Zero rRNA Removal Solution and Reaction Buffer at -80^{0}C
    d. Remove the reaction tubes and incubate at room temperature for 5 minutes.

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

5. Magnetic Bead Reaction and rRNA Removal
    a. Using a pipette, add the treated RNA from step *4* to the 1.5 mL tube containing the washed Magnetic Beads and immediately mix by pipetting at least 10 times. Then, vortex at medium setting for 10 seconds and place at room temperature
    b. Incubate the 1.5 mL tube at room temperature for 5 minutes
    c. Following incubation, mix the reations by vortexing at medium speed for 5 seconds and then place tube in heat block set to **50^0C for exactly 5 minutes.**
    d. Immediately transfer tube to magnetic stand and allow beads to separate for 5 minutes or until solution appears clear
    e. Carefully remove each supernatant (~90 uL) containing the RNA and transfer to a NEW RNase-free tube
        i. The supernatant contains the rRNA-depleted RNA
    f. Place the supernatant (RNA solution) on ice and immediately proceed to cleanup
6. **1.8X RNAclean** SPRI bead purification
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    b. In 1.5mL tube combine 90uL RNA + **162 uL beads**, immediately pipette mixture 10 times upon bead addition
    c. Close cap and incubate at room temperature for 20 minutes
    d. Separate beads on magnet for 5 minutes or until solution appears clear
    e. Remove and discard supernatant
    f. Keeping tubes on magnetic rack, wash the beads in 500 uL freshly prepared 75% Ethanol for at least 30 seconds
    g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
    h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
    i. Resuspend beads in **21 uL h20** and incubate at room temperature for 5 minutes
    j. Separate beads on magnet for 5 minutes or until solution appears clear
    k. Carefully transfer 18 uL Ribo-Zero'd RNA to new tube/plate
7. Proceed immediately to RNA fragmentation
8. *Recommended:* save 1 uL aliquot of Ribo-Zero'd RNA to check ribosomal RNA depletion using an Agilent Bioanalyzer RNA Pico 6000 kit

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

3. RNA FRAGMENTATION

Reagents:

1. 18 uL DNase-treated, rRNA-depleted RNA (cleaned up, in h20)
2. NEB Mg^{2+} Fragmentation Buffer
3. Agencourt RNAclean SPRI beads
4. *Freshly prepared* 75% Ethanol
5. (optional) Agilent Bioanalyzer RNA 6000 Pico Kit

Protocol:

Decontaminate work space, work quickly, and maintain RNase-free techniques

1. Combine in PCR-tube strip:
2. Incubate in PTC-225 DNA Engine Tetrad (MJ Research) thermocycler at 85^{0}C for 8 minutes (YES heated lid) and transfer immediately to ice

18 uL	DNase'd RNA
2 uL	NEB Mg^{2+} Frag Buffer
20 uL	

   a. 8 minutes yields fragments 100-1000 bp in length; 300 bp average. Increasing time will shift distribution towards smaller fragments.
   b. Place in thermocycler exactly as lid reaches 85^{0}C
   c. Remove and place on ice exactly when begins to lower temperature
   d. Proceed immediately to cleanup to STOP reaction
3. **1.8X RNAclean** SPRI bead purification
   a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
   b. Transfer 20 uL reaction to 96-well plate, add **36 uL beads**, and immediately pipette mixture 10 times upon bead addition
   c. Seal plate and incubate at room temperature for 20 minutes
   d. Separate beads on magnet for 5 minutes or until solution appears clear
   e. Remove and discard supernatant
   f. Keeping plate on magnetic rack, wash the beads in 200 uL freshly prepared 75% Ethanol for at least 30 seconds
   g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
   h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
   i. Resuspend beads in **9 uL h20** and incubate at room temperature for 5 minutes
   j. Separate beads on magnet for 5 minutes or until solution appears clear
   k. Carefully transfer 7 uL Fragmented RNA to new tube/plate
4. Proceed immediately to First-Strand Synthesis (FSS)
5. *Recommended:* save 1 uL aliquot of RNA to check fragmentation using an Agilent Bioanalyzer RNA Pico 6000 Kit

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

4. First Strand Synthesis (FSS)

Reagents:

1. 7 uL Fragmented RNA (cleaned-up, in h20)
2. Invitrogen 5X FS buffer
3. 3 ug/uL random primers ***see Special Notes section***
4. 10 mM dNTP mix
5. 100 mM DTT
6. Invitrogen Superscript III
7. Invitrogen RNaseOUT (40 U/uL)
8. Actinomycin D solution (1 ug/uL) ***see Special Notes section***
9. Agencourt RNAclean SPRI beads
10. Bio-Rad Micro Bio-Spin P-30 columns, Tris buffer (RNase-free)
11. *Freshly prepared* 75% Ethanol
12. (optional) Agilent Bioanalyzer RNA 6000 Pico Kit

Protocol:

1. Combine the following in a PCR tube-strip:

7 uL	Fragmented RNA
1 uL	3 ug/uL custom random primers
1 uL	10mM dNTP mix
9 uL	

2. Gently vortex and spin down
3. Incubate at 65^{0}C for 5 min in a PTC-225 DNA Engine Tetrad (MJ Research) thermocycler and transfer immediately to ice for 5 minutes
4. Add the following components to the reaction on ice:

4 uL	5X FS Buffer
1 uL	100 mM DTT
4 uL	1 ug/uL Actinomycin D solution
1 uL	RNaseOUT
1 uL	Superscript III (200 U)
11 uL	

5. Incubate in PTC-225 DNA Engine Tetrad (MJ Research) thermocycler:
    a. Set ramp speed for each transition to .1^{0}C/second, YES heated lid
        i. 5^{0}C – 5 minutes
        ii. 10^{0}C – 5 minutes
        iii. 15^{0}C – 5 minutes
        iv. 20^{0}C – 5 minutes

v. 25^{0}C – 5 minutes
vi. 30^{0}C – 5 minutes
vii. 35^{0}C – 5 minutes
viii. 42^{0}C – 30 minutes ← optional: @ 30" + 1uL SSIII (200 U)
ix. 45^{0}C – 10 minutes
x. 50^{0}C – 10 minutes
xi. 55^{0}C – 10 minutes
xii. Hold at 4^{0}C forever
xiii. **DO NOT HEAT KILL ENZYME**

6. Raise reaction volume to 50 uL by adding 30 uL h20
7. Clean-up with Micro Bio-Spin P-30 RNase-free column to remove dNTPs/ActD
    a. Obtain columns from 4^{0}C cold room (check that < 1yr old)
    b. Forcefully invert column 10 times whilst flicking to remove air bubbles
    c. Snap off fret and remove cap
    d. Let column drain for 3 minutes or until visibly drained
    e. Dump drained buffer and spin in supplied collection tube at 1000 x g for 2 minutes
    f. Place column in a new RNase-free collection tube (labeled appropriately)
    g. Load 50 uL sample to center of gel bed without disturbing it
    h. Spin at 1000 x g for 4 minutes (orient tubes as in previous spin)
    i. Sample is now in 10 mM Tris-HCl, pH 7.4
8. **1.8X RNAclean** SPRI bead purification to remove primers (+dNTPs/ActD)
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    b. Check that ~45 uL reaction remains in 1.5 mL tube, add **~80 uL beads**, and immediately pipette mixture 10 times upon bead addition
    c. Close cap and incubate at room temperature for 30 minutes
    d. Separate beads on magnet for 5 minutes or until solution appears clear
    e. Remove and discard supernatant
    f. Keeping plate on magnetic rack, wash the beads in 500 uL freshly prepared 75% Ethanol for at least 30 seconds
    g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
    h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
    i. Resuspend beads in **95 uL h20** and incubate at room temperature for 5 minutes
    j. Separate beads on magnet for 5 minutes or until solution appears clear
    k. Carefully transfer 88 uL RNA:cDNA hybrid to new tube/plate
9. Proceed immediately to Second-Strand Synthesis (SSS)
10. *Recommended:* save 1 uL aliquot of RNA:cDNA hybrid to check yield/size distribution/primer removal using an Agilent Bioanalyzer RNA Pico 6000 Kit
11. *Recommended:* save remainder of RNA:cDNA hybrid (~5 uL) for qPCR QC check
    a. Check genomic DNA removal using (–) enzyme control

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

5. Second Strand Synthesis

Reagents:

1. 88 uL RNA:cDNA hybrid (cleaned up 2X, in h20)
2. Invitrogen 5X FS Buffer
3. 100 mM DTT
4. Fermentas 2mM dACG-TP/4mM dU-TP mix
5. Invitrogen 5X SS Buffer
6. *E. coli* DNA ligase (10U/uL)
7. *E. coli* DNA polymerase (10U/uL)
8. *E. coli* RNase H (2U/uL)
9. 0.5 M EDTA
10. Agencourt AmpureXP SPRI beads
11. Qiagen Buffer EB
12. *Freshly prepared* 75% Ethanol
13. (optional) Agilent Bioanalyzer High Sensitivity DNA Kit

Protocol:

2. Combine the following in a PCR tube-strip:

88 uL	cDNA:RNA hybrid
4 uL	5X FS Buffer
2 uL	100 mM DTT
20 uL	2 mM DACG-TP/dU-TP mix
30 uL	5X SS Buffer
1 uL	*E. coli* DNA ligase (10 U)
4 uL	*E. coli* DNA polymerase (40 U)
1 uL	*E. coli* RNase H (2 U)
150 uL (+62 uL Master Mix)	

3. Mix gently by pipetting and incubate at **15^{0}C** in PTC-225 DNA Engine Tetrad (MJ Research) thermocycler for 2 hours
    a. **Do not vortex**
    b. NO HEATED LID Do not let sample temperature raise above 16^{0}C or DNA polymerase can strand displace
    c. Transfer immediately to ice
4. Add 10 uL 0.5M EDTA to STOP reaction and proceed immediately to cleanup
5. **1.8X AmpureXP** SPRI bead purification
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    b. Transfer ~155 uL reaction to 1.5mL tube, add **~280 uL beads**, and immediately pipette mixture 10 times upon bead addition
    c. Close cap and incubate at room temperature for 10 minutes

d. Separate beads on magnet for 5 minutes or until solution appears clear
e. Remove and discard supernatant
f. Keeping plate on magnetic rack, wash the beads in 500 uL freshly prepared 75% Ethanol for at least 30 seconds
g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
i. Resuspend beads in **88 uL Qiagen EB** and incubate at room temperature for 5 minutes
j. Separate beads on magnet for 5 minutes or until solution appears clear
k. Carefully transfer 85 uL dUTP-marked dsDNA to new tube/plate

6. ***SAFE STOPPING POINT***
    a. Proceed immediately to KAPA End Repair ***OR*** store dUTP-marked dsDNA at -20^{0}C for up to 1 week
7. *Recommended:* save 1 uL aliquot of dUTP-marked dsDNA to check yield/size distribution using an Agilent Bioanalyzer HS DNA chip

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

6. KAPA End Repair

Reagents:

1. 85 uL dUTP-marked dsDNA (cleaned-up, in EB)
2. KAPA library prep kit
3. Agencourt AmpureXP SPRI beads
4. Qiagen Buffer EB
5. *Freshly prepared* 75% Ethanol

Protocol:

1. Combine the following in a PCR tube strip:

85 uL	dUTP-marked dsDNA
10 uL	10X KAPA End Repair Buffer
5 uL	KAPA End Repair Enzyme Mix*
100 uL	

*15U T4 DNA Polymerase, 50U T4 Polynucleotide Kinase

2. Mix gently and incubate for 30 minutes at 20^{0}C in PTC-225 DNA Engine Tetrad (MJ Research) thermocycler
    a. No heated lid
    b. Transfer to ice and proceed immediately to cleanup
3. **1.8X AmpureXP** SPRI bead purification
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    b. Transfer 100uL reaction to 1.5mL tube, add **180 uL beads**, immediately pipette mixture 10 times upon bead addition
    c. Close cap and incubate at room temperature for 10 minutes
    d. Separate beads on magnet for 5 minutes or until solution appears clear
    e. Remove and discard supernatant
    f. Keeping tubes on magnetic rack, wash the beads in 500 uL freshly prepared 75% Ethanol for at least 30 seconds
    g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
    h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
    i. Resuspend beads in **33 uL Qiagen EB** and incubate at room temperature for 5 minutes
    j. Separate beads on magnet for 5 minutes or until solution appears clear
    k. Carefully transfer 31 uL End-Repaired dsDNA to new tube/plate
4. ***SAFE STOPPING POINT***
    a. Proceed immediately to KAPA A-Tailing ***OR*** store End-Repaired dsDNA at -20^{0}C for up to 1 week

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

7. KAPA A-Tailing

Reagents:

1. 30 uL End-Repaired dsDNA (cleaned-up, in EB)
2. KAPA library prep kit
3. Agencourt AmpureXP SPRI beads
4. Qiagen Buffer EB
5. *Freshly prepared* 75% Ethanol

Protocol:

1. Combine the following in a PCR tube strip:

30 uL	End-Repaired dsDNA
12 uL	H_20
5 uL	10X KAPA A-Tailing Buffer*
3 uL	KAPA A-Tailing Enzyme**
50 uL	

*contains 2 mM dATP (.2 mM final)

**15U 3'→5' exo$^-$ klenow fragment

2. Mix gently and incubate for 30 minutes at 30^{0}C in PTC-225 DNA Engine Tetrad (MJ Research) thermocycler
    a. No heated lid
    b. Transfer to ice and proceed immediately to cleanup
3. **1.8X AmpureXP** SPRI bead purification
    b. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    c. Add **90 uL beads**, immediately pipette mixture 10 times upon bead addition
    d. Close cap and incubate at room temperature for 10 minutes
    e. Separate beads on magnet for 5 minutes or until solution appears clear
    f. Remove and discard supernatant
    g. Keeping tubes on magnetic rack, wash the beads in 500 uL freshly prepared 75% Ethanol for at least 30 seconds
    h. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
    i. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
    j. Resuspend beads in **33.5 uL Qiagen EB and** incubate at room temperature for 5 minutes
    k. Separate beads on magnet for 5 minutes or until solution appears clear
    l. Carefully transfer 31 uL A-Tailed dsDNA to new tube/plate
4. ***SAFE STOPPING POINT***
    m. Proceed immediately to KAPA Adapter Ligation ***OR*** store A-Tailed dsDNA at -20^{0}C for up to 1 week

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

8. KAPA Adapter Ligation

Reagents:

1. 30 uL A-Tailed dsDNA (cleaned-up, in EB)
2. KAPA library prep kit
3. 15 uM Barcoded Y-Adapters
4. Agencourt AmpureXP SPRI beads
5. Qiagen Buffer EB
6. *Freshly prepared* 75% Ethanol

Protocol:

Change gloves/clean pipette every time handle different adapter

1. Dilute adapters approximately 1:10 in h20 (1.5 uM final). See calculation below.
2. Combine the following in a PCR tube strip:

30 uL	A-Tailed dsDNA
10 uL	5X KAPA Ligation Buffer*
5 uL	1.5 uM Barcoded Y-Adapter #_____
5 uL	KAPA Ligation Enzyme Mix**
50 uL	

*Contains 30% PEG 6000 (6% final). Raising PEG concentration lowers SPRI bead size selection (i.e. smaller fragments elute off beads than normal)

**30,000U T4 DNA Ligase

***Calculate molar ratio of dsDNA target:Adapter to ensure in the 1:15 regime. Remember that each dsDNA fragment represents 2 targets for adapter ligation but that ER/A-tailing is not 100% efficient…. If assume 50% efficiency these scaling factors cancel each other out. Example calculation:

(30 ng A-Tailed dsDNA, 300 bp)*(1 mole/300bp*650g) = $1.54 * 10^{-4}$ nmoles
(5uL * 1.5 uM Adapters)*(L/10^6 uL) = $2.5 * 10^{-3}$ nmoles
~1:15

3. Mix gently and incubate for 15 minutes at 20^{0}C in PTC-225 DNA Engine Tetrad (MJ Research) thermocycler
    n. No heated lid
    o. Transfer immediately to ice
4. Proceed immediately to SPRI Bead Adapter removal

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

9. SPRI Bead Adapter Removal

Reagents:

1. 50 uL Adapter-Ligated dsDNA
2. Agencourt AmpureXP SPRI beads
3. Qiagen Buffer EB
4. *Freshly prepared* 75% Ethanol
5. (optional) Agilent Bioanalyzer High Sensitivity DNA Kit

Protocol:

1. **1.0X AmpureXP** SPRI bead purification
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles
    b. Add **50 uL beads**, immediately pipette mixture 10 times upon bead addition
    c. Close cap and incubate at room temperature for 15 minutes
    d. Separate beads on magnet for 5 minutes or until solution appears clear
    e. Remove and discard supernatant
    f. Keeping tubes on magnetic rack, wash the beads in 500 uL freshly prepared 75% Ethanol for at least 30 seconds
    g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
    h. Allow beads to air dry for 15 minutes (ensure no ethanol traces remain)
    i. Resuspend beads in **22 uL Qiagen EB and** incubate at room temperature for 5 minutes
    j. Separate beads on magnet for 5 minutes or until solution appears clear
    k. Carefully transfer 20 uL Adapter-Ligated dsDNA to new tube/plate
2. ***SAFE STOPPING POINT***
    a. Proceed immediately to USER digestion ***OR*** store Adapter-Ligated dsDNA at -20^{0}C for up to 1 week
3. *Recommended:* save 1 uL aliquot of Adapter-Ligated dsDNA to check yield/size distribution/adapter dimer contamination using an Agilent Bioanalyzer HS DNA chip.

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

10. USER digestion (DO NOT FORGET TO DIGEST dUTP-marked strand)

Reagents:

1. 47 uL Adapter-ligated dsDNA (cleaned up 1X or 2X)
2. NEB USER (Uracil-Specific Excision Reagent) enzyme (1U/uL)
3. (optional) Agilent Bioanalyzer RNA Pico Kit
4. (optional) Agilent Bioanalyzer High Sensitivity DNA Kit

Protocol:

1. Add 1 uL USER enzyme to each reaction
2. Incubate at 37^{0}C for 30 minutes, followed by 25^{0}C for 15 minutes
    a. USER is not heat inactivated
    b. Yes heated lid
3. ***SAFE STOPPING POINT***
    a. Proceed immediately to KAPA Real-Time PCR Amplification ***OR*** store USER-digested ssDNA at -20^{0}C
4. *Recommended:* save 1 uL aliquot of USER-digested ssDNA to check yield/size distribution/adapter dimer contamination using an Agilent Bioanalyzer RNA Pico 6000 Kit
5. *Recommended:* save 1 uL aliquot of USER-digested ssDNA to check second strand digestion using an Agilent Bioanalyzer HS DNA chip

Supplementary Protocol 3.1 (Continued) | Strand-specific, non-polyA-selected *P. falciparum* library preparation protocol

11. KAPA Real-Time PCR Amplification

Reagents:

1. 9 uL USER-digested ssDNA (~1/5 reaction used as input)
2. KAPA real-time library amplification kit
3. 12.5 uM *each* Forward and Reverse Indexing PCR Primer Mix
4. Agencourt AmpureXP SPRI beads
5. Qiagen Buffer EB
6. *Freshly prepared* 75% Ethanol
7. Agilent Bioanalyzer High Sensitivity DNA Kit

Protocol:

1. Add 65uL of fluorescence references 1-4 to a 96-well plate
2. Combine the following for each library in 96-well plate:

9 uL	ssDNA sample (USER-digested)
2 uL	F&R Indexing PCR Primer Mix*
11 uL	2X KAPA HiFi SYBR/Rox qPCR Mix
22 uL	

*1 uM each (final)

3. Seal, vortex and spin down plate
4. Transfer 3 X 20 uL of each reference to a 384-well plate using multi-channel pipette
5. Add 20 uL of each library to 384-well plate
6. Seal and spin down plate
    a. Ensure no debris or air bubbles
7. Thermocycle on ABI 7900 Real-Time PCR Machine with following program
    a. 98^{0}C – 2 minutes
    b. 20 cycles:
        i. 98^{0}C – 20 seconds
        ii. 55^{0}C* – 180 seconds
            *Optimal annealing at 55^{0}C for BROAD indexing PCR primers
        iii. 55^{0}C – 10 seconds **DATA ACQUISITION**
8. SIT IN ROOM AND WATCH THE PCR MACHINE THE ENTIRE TIME TO STOP REACTION WHEN LIBRARY BETWEEN REFERENCE 1 and 3
    a. Must stop during data acquisition phase before machine ramps to 98^{0}C
    b. The less PCR cycles the better. Do not over-cycle.
    c. If expect vastly different yields, amplify libraries individually or perform a diagnostic amplification and repeat with normalized sample input.
9. **1.0X AmpureXP** SPRI bead purification
    a. Equilibrate beads to room temperature for 30 minutes, vortex bead slurry thoroughly before adding to RNA and pipette beads slowly to avoid forming air bubbles

b. Transfer 20uL reaction to 96-well plate, add 20uL beads, immediately pipette mixture 10 times upon bead addition
c. Seal plate and incubate at room temperature for 10 minutes
d. Separate beads on magnet for 5 minutes or until solution appears clear
e. Remove and discard supernatant
f. Keeping plate on magnetic rack, wash the beads in 200 uL freshly prepared 75% Ethanol for at least 30 seconds
g. Carefully remove and discard ethanol without disturbing pellet and repeat the process for a total of **2 washes**
h. Allow beads to air dry for 10 minutes (ensure no ethanol traces remain)
i. Resuspend beads in **25 uL Qiagen EB and** incubate at room temperature for 5 minutes
j. Separate beads on magnet for 5 minutes or until solution appears clear
k. Carefully transfer 22.5 uL PCR-amplified dsDNA to new tube/plate

6. ***SAFE STOPPING POINT***
   a. Proceed immediately to KAPA Library Quantification ***OR*** store PCR-amplified dsDNA at -20^{0}C
7. *Recommended:* save 1 uL aliquot of PCR-amplified final library to check yield/size distribution/adapter dimer contamination using an Agilent Bioanalyzer HS DNA chip. If necessary, repeat 1.0X AmpureXP SPRI bead purification to further remove adapter dimers.

Special Notes

*Ideally, custom order 76% AT-biased random hexamers to reflect the approximate AT balance of the *P. falciparum* transcriptome.

*Prepare the Actinomycin D solution freshly in water, keep it protected from light, and handle it with caution, as ActD is highly toxic. Verify the concentration of ActD solution using Beer's law immediately prior to use. Chilling the ActD solution may increase solubility (!).

Supplementary Table 3.1 | Sequencing metrics.

Sample ID	Lane	Lane type	Read length	Total reads	Pass filter (PF) reads	PF reads %	Strand balance %
R	1	Paired	101	36,309,330	34,716,924	95.61	50
ET	1	Paired	101	33,627,496	32,208,510	95.78	50
LT	1	Paired	101	34,634,574	33,184,560	95.81	49.98
S	1	Paired	101	36,181,466	34,686,070	95.87	49.96
T6	1	Paired	101	35,928,966	34,401,300	95.75	50.02
T14	1	Paired	101	40,993,176	39,174,956	95.56	50.01
T20	1	Paired	101	41,758,410	39,868,948	95.48	50.01
T24	2	Paired	101	41,510,498	38,772,456	93.4	50.02
T28	2	Paired	101	49,140,432	45,863,464	93.33	50
T32	2	Paired	101	43,502,870	40,844,278	93.89	49.98
T36	2	Paired	101	48,691,486	45,536,284	93.52	49.99
T40	2	Paired	101	44,405,570	41,662,908	93.82	49.98
T44	2	Paired	101	44,358,344	41,491,020	93.54	49.98
T48	2	Paired	101	40,561,706	37,815,894	93.23	49.99
TT8	2	Paired	101	41,977,026	38,868,142	92.59	50

Strand balance %	Read pair duplicates	Duplicate %	Estimated library size	Median insert size	Mean insert size	Adapter %
50	5,995,382	35.12	18,865,157	205	233	0.0031
50	4,803,198	30.49	21,127,976	213	243	0.003
49.98	5,645,437	35.03	17,904,375	210	239	0.0073
49.96	6,863,514	40.54	15,201,954	207	235	0.0109
50.02	9,398,673	57.61	8,052,612	201	224	0.1125
50.01	5,644,712	32.35	21,713,241	196	222	0.0877
50.01	5,263,531	28.23	28,001,255	206	232	0.0088
50.02	5,788,680	30.34	25,968,555	208	233	0.0032
50	5,536,468	24.65	40,272,749	206	234	0.0006
49.98	5,672,423	28.61	29,489,373	206	235	0.0014
49.99	5,634,662	25.13	39,166,904	207	234	0.0004
49.98	5,999,535	29.15	29,527,040	205	233	0.0004
49.98	6,405,193	31.19	26,937,480	210	238	0.0007
49.99	8,133,834	42.85	15,676,773	214	241	0.0013
50	5,321,991	34.06	18,135,700	199	225	0.0101

Supplementary Table 3.2 | Aligned read metrics.

Sample ID	Uniquely aligned reads	Aligned reads %	Reads aligned to sense %	rRNA %	Mean insert size	rRNA-filtered mean GC %
R	23,076,722	63.56	99.32	13.44	205	27.59
ET	22,241,738	66.14	99.50	10.94	212	27.42
LT	22,537,304	65.07	99.07	10.14	209	28.10
S	23,455,234	64.83	99.14	12.82	208	27.60
T6	22,038,452	61.34	99.28	6.55	203	25.38
T14	24,359,556	59.42	99.19	4.88	203	25.74
T20	25,137,560	60.20	99.26	9.37	206	24.95
T24	25,185,524	60.67	99.48	12.72	209	24.91
T28	30,428,074	61.92	99.40	6.33	208	24.85
T32	26,377,500	60.63	99.28	9.83	206	24.66
T36	30,713,000	63.08	99.42	7.56	209	24.57
T40	27,681,836	62.34	99.42	10.11	208	24.90
T44	27,825,036	62.73	99.44	5.62	207	25.02
T48	24,889,826	61.36	99.81	18.32	210	26.15
TT8	20,498,156	48.83	98.92	3.26	201	25.31

Exonic %	Intronic %	Intergenic %	Top 2000 genes CV	Top 2000 genes gap %	Top 1000 genes CV	Top 4000 genes CV
87.30	6.90	5.80	0.26	0.1	0.24	0.33
89.60	6.70	3.60	0.32	0.1	0.28	0.39
86.60	8.90	4.50	0.32	0.1	0.29	0.41
89.50	6.70	3.70	0.33	0.1	0.31	0.39
87.50	6.00	6.50	0.29	0.4	0.26	0.33
83.50	8.60	7.80	0.27	0.1	0.25	0.33
82.40	8.60	9.00	0.24	0.1	0.23	0.32
86.60	6.90	6.60	0.25	0.1	0.24	0.31
84.70	7.40	7.90	0.23	0	0.23	0.26
83.50	9.10	7.40	0.25	0.1	0.24	0.28
86.60	7.10	6.30	0.25	0.1	0.24	0.27
87.00	6.90	6.20	0.25	0.1	0.24	0.28
86.20	6.90	6.90	0.27	0.2	0.26	0.3
90.90	5.20	3.90	0.31	0.5	0.29	0.36
82.50	8.10	9.40	0.27	0.1	0.26	0.33

Supplementary Table 3.3 | GO term enrichment results.

Ring-specific GO term	Ring-specific GO term p-value
cell-cell adhesion	4.95E-18
rosetting	1.93E-17
heterophilic cell adhesion	1.93E-17
adhesion to other organism during symbiotic interaction	3.16E-16
cytoadherence to microvasculature, mediated by parasite protein	3.16E-16
adhesion to host	3.16E-16
cell adhesion	4.42E-16
biological adhesion	6.31E-15
pathogenesis	4.88E-14
interaction with host	5.28E-14
gene expression	2.17E-11
translation	3.14E-05
RNA processing	0.000144
macromolecule biosynthetic process	0.000146
RNA metabolic process	0.00017
ribosome biogenesis and assembly	0.000498

Trophozoite-specific GO term	Trophozoite-specific GO term p-value
DNA replication	0.00105
amino acid and derivative metabolic process	0.00105
nitrogen compound metabolic process	0.00105
carboxylic acid metabolic process	0.00105
organic acid metabolic process	0.00105
amine metabolic process	0.00105
amino acid metabolic process	0.00124
biosynthetic process	0.00164
protein folding	0.00342
aldehyde metabolic process	0.00462

Supplementary Table 3.3 (Continued) | GO term enrichment results.

Schizont-specific GO term	Schizont-specific GO term p-value
proteolysis	0.000000348
DNA metabolic process	0.0000119
DNA replication initiation	0.0000121
protein catabolic process	0.0000234
ubiquitin-dependent protein catabolic process	0.0000234
post-translational protein modification	0.0000538
biopolymer catabolic process	0.0000538
modification-dependent protein catabolic process	0.0000593
modification-dependent macromolecule catabolic process	0.0000593
chromosome organization and biogenesis	0.000191
chromatin assembly or disassembly	0.000202
cellular protein catabolic process	0.000372
proteolysis involved in cellular protein catabolic process	0.000372
DNA packaging	0.000372
tricarboxylic acid cycle	0.000668
cellular respiration	0.000668
aerobic respiration	0.000668
coenzyme catabolic process	0.000668
energy derivation by oxidation of organic compounds	0.000668
acetyl-CoA catabolic process	0.000668
cofactor catabolic process	0.000668
establishment and/or maintenance of chromatin structure	0.000911
cytoskeleton organization and biogenesis	0.00152
chromatin assembly	0.00157
acetyl-CoA metabolic process	0.00157
localization	0.00183
nucleosome organization	0.00183
protein modification process	0.00183
biopolymer modification	0.00218
transport	0.00218

Supplementary Table 3.4 | Cufflinks assembled transcript property metrics.

	Cufflinks transcripts (excluding un-annotated)	**Cufflinks RABT transcripts (excluding un-annotated)**	***PlasmoDBv10.0* annotated transcripts**
Average Maximum FPKM	165	238	469
Average average FPKM	57	83	164
Average length	3223	2618	2197
Average GC content	20.8%	24.0%	25.4%
Single exon rate	46.6%	45.0%	47.8%
Maximum exon count	32	34	34

Supplementary Table 3.5 | Putative *P. falciparum* circRNAs with at least five unique supporting reads.

Chromosome	Start	Stop	Supporting reads	Strand	Unique Supporting reads	Length (genomic distance)	Associated annotation (if applicable)
Pf3D7_09_v3	913503	913614	216	+	123	111	PF3D7_092250
Pf3D7_11_v3	222005	222116	78	-	65	111	PF3D7_110500
Pf3D7_09_v3	421099	421330	71	-	56	231	PF3D7_090930
Pf3D7_14_v3	323890	324002	53	+	51	112	PF3D7_140860
Pf3D7_14_v3	1103004	1103104	72	+	51	100	PF3D7_142820
Pf3D7_11_v3	222005	222107	62	-	51	102	PF3D7_110500
Pf3D7_03_v3	315293	315518	63	-	51	225	PF3D7_030720
Pf3D7_08_v3	687808	687919	53	+	48	111	PF3D7_081420
Pf3D7_09_v3	270975	271110	58	-	46	135	PF3D7_090540
Pf3D7_10_v3	812336	812553	59	+	44	217	PF3D7_102000
Pf3D7_06_v3	1033279	1033401	45	-	41	122	PF3D7_062540
Pf3D7_14_v3	3036471	3037909	41	-	35	1438	PF3D7_147420
Pf3D7_06_v3	895390	895570	35	-	34	180	PF3D7_062190
Pf3D7_05_v3	1071423	1071531	34	+	31	108	PF3D7_052580
Pf3D7_02_v3	104059	104221	33	-	31	162	PF3D7_020200
Pf3D7_14_v3	2711332	2711437	33	+	30	105	PF3D7_146630
Pf3D7_06_v3	748384	748495	36	-	30	111	PF3D7_061780
Pf3D7_02_v3	306192	306309	35	-	29	117	PF3D7_020760
Pf3D7_06_v3	795729	795936	27	-	26	207	PF3D7_061900
Pf3D7_12_v3	162386	162496	30	-	26	110	PF3D7_120290
Pf3D7_13_v3	2187585	2187690	27	+	23	105	PF3D7_135510
Pf3D7_10_v3	1326738	1326897	26	-	23	159	PF3D7_103310
Pf3D7_14_v3	2455290	2455394	25	+	23	104	PF3D7_146010
Pf3D7_07_v3	717525	717677	36	+	22	152	PF3D7_071630
Pf3D7_10_v3	1425378	1425486	25	+	22	108	PF3D7_103590
Pf3D7_01_v3	342994	343120	21	+	20	126	PF3D7_010830
Pf3D7_05_v3	64833	64959	20	-	20	126	PF3D7_050120
Pf3D7_14_v3	1548535	1549065	23	+	20	530	PF3D7_143840
Pf3D7_10_v3	1421568	1421679	21	+	19	111	PF3D7_103580
Pf3D7_10_v3	1098805	1100490	21	-	19	1685	PF3D7_102600
Pf3D7_10_v3	1425330	1425486	21	+	19	156	PF3D7_103590
Pf3D7_02_v3	207222	207364	18	-	18	142	PF3D7_020470
Pf3D7_10_v3	341449	341557	21	+	18	108	PF3D7_100810
Pf3D7_04_v3	386297	386411	19	+	18	114	PF3D7_040780
Pf3D7_09_v3	1458815	1460208	21	+	18	1393	PF3D7_093680
Pf3D7_11_v3	235878	235982	18	-	18	104	PF3D7_110540
Pf3D7_14_v3	2825818	2825923	19	+	17	105	PF3D7_146870
Pf3D7_12_v3	148983	149369	20	-	16	386	PF3D7_050590
Pf3D7_02_v3	306204	306309	21	-	16	105	PF3D7_020760
Pf3D7_04_v3	666592	667254	19	+	16	662	PF3D7_041490
Pf3D7_14_v3	1016314	1016599	16	-	16	285	PF3D7_142600
Pf3D7_14_v3	799419	799576	18	-	16	157	PF3D7_141920
Pf3D7_10_v3	165230	165338	18	-	15	108	PF3D7_100360
Pf3D7_14_v3	323887	324002	15	+	15	115	PF3D7_140860
Pf3D7_11_v3	226728	226961	15	+	14	233	PF3D7_142790
Pf3D7_14_v3	1247670	1247938	14	+	14	268	PF3D7_143170
Pf3D7_14_v3	782077	782203	14	-	14	126	PF3D7_141880
Pf3D7_05_v3	713249	713364	16	-	14	115	PF3D7_051690
Pf3D7_09_v3	270825	271121	13	+	13	296	PF3D7_090540
Pf3D7_11_v3	815567	816356	23	-	13	789	PF3D7_112160
Pf3D7_13_v3	2052422	2052601	13	+	13	179	PF3D7_135140
Pf3D7_09_v3	1459118	1459313	15	+	13	195	PF3D7_093680
Pf3D7_13_v3	911758	911902	13	-	13	144	PF3D7_132190
Pf3D7_08_v3	687808	687934	13	+	12	126	PF3D7_081420

Supplementary Table 3.5 (Continued) | Putative *P. falciparum* circRNAs with at least five unique supporting reads.

Pf3D7_14_v3	2091862	2091985	14	+	12	123	PF3D7_145110
Pf3D7_05_v3	1057813	1057951	16	+	12	138	PF3D7_052540
Pf3D7_11_v3	950736	950896	14	+	12	160	NA
Pf3D7_06_v3	524430	525770	14	-	12	1340	PF3D7_061260
Pf3D7_07_v3	137323	137443	15	+	12	120	PF3D7_070350
Pf3D7_05_v3	317982	318099	12	-	12	117	PF3D7_050770
Pf3D7_04_v3	1059219	1059324	12	+	12	105	PF3D7_042350
Pf3D7_09_v3	1203756	1203881	11	+	11	125	PF3D7_093030
Pf3D7_09_v3	1204125	1204314	14	-	11	189	PF3D7_093030
Pf3D7_07_v3	277949	278054	13	-	11	105	PF3D7_070550
Pf3D7_08_v3	811462	812303	11	-	11	841	PF3D7_081770
Pf3D7_02_v3	766505	766638	13	-	11	133	PF3D7_021880
Pf3D7_08_v3	687961	688066	13	+	11	105	PF3D7_081420
Pf3D7_03_v3	706209	706371	11	-	10	162	PF3D7_031730
Pf3D7_09_v3	322819	322928	10	-	10	109	PF3D7_090660
Pf3D7_08_v3	433685	433822	10	+	10	137	PF3D7_080860
Pf3D7_04_v3	93202	93355	11	-	10	153	PF3D7_040160
Pf3D7_14_v3	1247677	1247778	12	+	10	101	PF3D7_143170
Pf3D7_09_v3	157500	157618	10	-	10	118	PF3D7_090340
Pf3D7_11_v3	1430964	1437008	10	-	10	6044	PF3D7_113650
Pf3D7_11_v3	353243	353363	10	+	10	120	PF3D7_110800
Pf3D7_10_v3	1141172	1141286	10	-	10	114	PF3D7_102730
Pf3D7_10_v3	107668	107776	13	-	10	108	PF3D7_100210
Pf3D7_13_v3	2062151	2062268	10	+	10	117	PF3D7_135170
Pf3D7_11_v3	1495268	1495406	13	+	10	138	PF3D7_113800
Pf3D7_07_v3	1330733	1330964	10	+	10	231	PF3D7_073090
Pf3D7_10_v3	81226	81543	11	-	10	317	PF3D7_100150
Pf3D7_05_v3	64815	64959	10	-	10	144	PF3D7_050120
Pf3D7_10_v3	1457144	1457246	11	-	10	102	PF3D7_103690
Pf3D7_13_v3	632574	633169	10	+	10	595	PF3D7_131490
Pf3D7_11_v3	1433155	1437008	10	-	10	3853	PF3D7_113650
Pf3D7_04_v3	762504	762606	11	+	10	102	PF3D7_041740
Pf3D7_09_v3	374890	375007	12	-	10	117	PF3D7_090810
Pf3D7_11_v3	1400498	1400631	10	-	10	133	PF3D7_113590
Pf3D7_14_v3	1995270	1995378	11	-	10	108	PF3D7_144850
Pf3D7_10_v3	1231525	1231627	12	-	10	102	PF3D7_103020
Pf3D7_04_v3	104555	104669	10	+	9	114	PF3D7_040180
Pf3D7_05_v3	64851	64959	10	-	9	108	PF3D7_050120
Pf3D7_08_v3	657346	657454	9	+	9	108	PF3D7_081330
Pf3D7_11_v3	543076	543599	12	-	9	523	PF3D7_111420
Pf3D7_06_v3	1172135	1173037	9	-	9	902	PF3D7_062830
Pf3D7_01_v3	414916	415025	9	-	9	109	PF3D7_011070
Pf3D7_14_v3	1844801	1844930	9	-	9	129	PF3D7_144480
Pf3D7_06_v3	748516	748651	9	-	9	135	PF3D7_061780
Pf3D7_09_v3	1427434	1427592	13	+	9	158	NA
Pf3D7_04_v3	104555	104729	9	+	9	174	PF3D7_040180
Pf3D7_06_v3	1123636	1123741	19	-	9	105	PF3D7_062810
Pf3D7_09_v3	1176736	1176913	9	+	9	177	PF3D7_092940
Pf3D7_08_v3	687808	687961	11	+	9	153	PF3D7_081420
Pf3D7_05_v3	715867	715979	9	-	9	112	PF3D7_051700
Pf3D7_08_v3	1207128	1207236	10	-	9	108	PF3D7_082790
Pf3D7_08_v3	845195	847517	11	-	9	2322	PF3D7_081860
Pf3D7_13_v3	878362	878519	8	-	8	157	PF3D7_132130
Pf3D7_11_v3	1961815	1961941	8	+	8	126	PF3D7_120260
Pf3D7_06_v3	1171201	1171973	9	-	8	772	PF3D7_062830
Pf3D7_07_v3	339644	339767	8	+	8	123	PF3D7_070730
Pf3D7_13_v3	151216	151342	8	-	8	126	PF3D7_130280
Pf3D7_07_v3	274413	274527	8	-	8	114	PF3D7_070550

Supplementary Table 3.5 (Continued) | Putative *P. falciparum* circRNAs with at least five unique supporting reads.

Pf3D7_08_v3	687919	688054	8	+	8	135	PF3D7_081420	
Pf3D7_02_v3	800822	800930	8	+	8	108	PF3D7_022000	
Pf3D7_06_v3	768777	768956	8	+	8	179	PF3D7_061830	
Pf3D7_11_v3	139852	139989	9	-	8	137	NA	
Pf3D7_11_v3	773065	773223	8	+	8	158	PF3D7_112050	
Pf3D7_04_v3	93066	93180	8	-	8	114	PF3D7_040160	
Pf3D7_05_v3	1123122	1123254	8	+	8	132	NA	
Pf3D7_14_v3	1320487	1320610	8	-	8	123	PF3D7_143340	
Pf3D7_07_v3	382721	382906	9	+	8	185	PF3D7_070840	
Pf3D7_12_v3	670843	670960	10	+	8	117	PF3D7_121690	
Pf3D7_05_v3	1308357	1308478	8	+	8	121	PF3D7_053230	
Pf3D7_05_v3	295120	295332	8	-	8	212	PF3D7_050710	
Pf3D7_04_v3	632535	632670	7	-	7	135	PF3D7_041400	
Pf3D7_08_v3	567941	568184	7	+	7	243	PF3D7_081130	
Pf3D7_04_v3	871555	871666	7	+	7	111	PF3D7_041960	
Pf3D7_05_v3	1082144	1082246	7	-	7	102	PF3D7_052620	
Pf3D7_08_v3	1304355	1304457	7	+	7	102	PF3D7_083060	
Pf3D7_08_v3	687808	688033	7	+	7	225	PF3D7_081420	
Pf3D7_09_v3	473244	473363	8	+	7	119	PF3D7_091020	
Pf3D7_05_v3	84561	84669	9	-	7	108	PF3D7_050160	
Pf3D7_12_v3	928261	928434	8	+	7	173	PF3D7_122310	
Pf3D7_05_v3	758663	758790	7	+	7	127	PF3D7_051820	
Pf3D7_14_v3	289161	289269	8	+	7	108	PF3D7_140790	
Pf3D7_08_v3	1252862	1253625	7	+	7	763	PF3D7_082910	
Pf3D7_09_v3	241930	242134	7	+	7	204	PF3D7_090510	
Pf3D7_10_v3	727098	727245	9	-	7	147	PF3D7_101820	
Pf3D7_08_v3	687808	687949	7	+	7	141	PF3D7_081420	
Pf3D7_12_v3	831841	831952	7	-	7	111	PF3D7_122090	
Pf3D7_10_v3	114807	114915	7	+	7	108	PF3D7_100220	
Pf3D7_02_v3	278039	278144	8	-	7	105	PF3D7_020700	
Pf3D7_06_v3	1172135	1173721	8	-	7	1586	PF3D7_062830	
Pf3D7_09_v3	653583	653694	7	-	7	111	PF3D7_091540	
Pf3D7_12_v3	73695	73815	9	-	7	120	PF3D7_120090	
Pf3D7_11_v3	1249394	1249565	7	-	7	171	PF3D7_113220	
Pf3D7_02_v3	252756	253484	7	+	7	728	PF3D7_020630	
Pf3D7_13_v3	1137874	1137980	7	-	7	106	PF3D7_132700	
Pf3D7_13_v3	2477303	2477438	8	-	7	135	PF3D7_136180	
Pf3D7_06_v3	454379	454555	7	+	7	176	PF3D7_061080	
Pf3D7_10_v3	1400988	1401168	7	+	7	180	PF3D7_103530	
Pf3D7_03_v3	524921	525211	7	-	7	290	PF3D7_031240	
Pf3D7_13_v3	2616125	2618062	8	-	7	1937	PF3D7_136540	
Pf3D7_07_v3	759540	760248	11	-	7	708	PF3D7_071760	
Pf3D7_13_v3	2435964	2436078	7	+	7	114	PF3D7_136080	
Pf3D7_13_v3	2120087	2120189	7	-	7	102	PF3D7_135300	
Pf3D7_04_v3	104555	104699	7	+	7	144	PF3D7_040180	
Pf3D7_10_v3	1168067	1168187	9	+	7	120	PF3D7_102830	
Pf3D7_02_v3	349130	349232	7	-	7	102	PF3D7_020840	
Pf3D7_13_v3	1653743	1653968	8	-	7	225	PF3D7_134200	
Pf3D7_11_v3	932157	932433	7	-	7	276	PF3D7_112350	
Pf3D7_05_v3	254464	255559	7	-	7	1095	PF3D7_050590	
Pf3D7_13_v3	953152	954054	7	+	7	902	PF3D7_132250	
Pf3D7_04_v3	93202	93324	6	-	6	122	PF3D7_040160	
Pf3D7_05_v3	1226931	1227131	6	+	6	200	PF3D7_053010	
Pf3D7_07_v3	613558	613676	6	-	6	118	PF3D7_063250	
Pf3D7_09_v3	1135907	1136188	6	+	6	281	PF3D7_092800	
Pf3D7_08_v3	149146	149350	6	+	6	204	PF3D7_080200	
Pf3D7_06_v3	1172135	1173355	6	-	6	1220	PF3D7_062830	

Supplementary Table 3.5 (Continued) | Putative *P. falciparum* circRNAs with at least five unique supporting reads.

Pf3D7_13_v3	1419327	1419564	7	-	6	237	PF3D7_133510	
Pf3D7_12_v3	989084	989215	6	+	6	131	PF3D7_122430	
Pf3D7_02_v3	105049	105172	6	-	6	123	PF3D7_020200	
Pf3D7_10_v3	449519	449744	6	+	6	225	PF3D7_101170	
Pf3D7_07_v3	858926	859031	6	-	6	105	PF3D7_071960	
Pf3D7_09_v3	1206088	1206203	6	+	6	115	PF3D7_093030	
Pf3D7_14_v3	1092998	1093117	6	+	6	119	PF3D7_142790	
Pf3D7_04_v3	192469	192595	7	-	6	126	PF3D7_040340	
Pf3D7_10_v3	475424	475589	6	-	6	165	PF3D7_101240	
Pf3D7_10_v3	122253	122993	7	+	6	740	PF3D7_100240	
Pf3D7_10_v3	920722	920827	6	+	6	105	PF3D7_102200	
Pf3D7_12_v3	298589	298700	8	-	6	111	PF3D7_120660	
Pf3D7_13_v3	632574	633492	6	+	6	918	PF3D7_131490	
Pf3D7_14_v3	2604524	2605189	7	-	6	665	PF3D7_146440	
Pf3D7_11_v3	1966634	1966760	6	-	6	126	PF3D7_114900	
Pf3D7_07_v3	632766	632915	6	-	6	149	PF3D7_071390	
Pf3D7_05_v3	876055	876159	6	+	6	104	PF3D7_052140	
Pf3D7_06_v3	257563	259970	7	+	6	2407	PF3D7_060610	
Pf3D7_14_v3	2392771	2392879	6	+	6	108	PF3D7_145830	
Pf3D7_05_v3	325677	325785	7	+	6	108	PF3D7_050780	
Pf3D7_03_v3	315573	315983	7	-	6	410	PF3D7_030720	
Pf3D7_13_v3	418777	419424	7	+	6	647	PF3D7_130910	
Pf3D7_04_v3	565305	565445	6	-	6	140	PF3D7_041270	
Pf3D7_10_v3	199455	199585	6	-	6	130	PF3D7_100430	
Pf3D7_09_v3	740226	740347	6	-	6	121	PF3D7_091790	
Pf3D7_04_v3	501141	501249	6	-	6	108	PF3D7_041100	
Pf3D7_05_v3	1037806	1037910	6	+	6	104	PF3D7_052500	
Pf3D7_08_v3	687838	687985	6	+	6	147	PF3D7_081420	
Pf3D7_12_v3	570208	570366	7	+	6	158	PF3D7_121290	
Pf3D7_14_v3	1911799	1911937	6	+	6	138	PF3D7_144650	
Pf3D7_12_v3	726958	727093	7	-	6	135	PF3D7_121850	
Pf3D7_13_v3	43904	44073	7	-	6	169	PF3D7_130030	
Pf3D7_11_v3	637022	637178	7	-	6	156	PF3D7_111680	
Pf3D7_09_v3	154185	154335	6	+	6	150	PF3D7_090340	
Pf3D7_13_v3	2242116	2242263	6	-	6	147	PF3D7_135660	
Pf3D7_09_v3	1051312	1051423	6	-	6	111	PF3D7_092610	
Pf3D7_10_v3	998424	998527	6	-	6	103	PF3D7_102390	
Pf3D7_07_v3	1350397	1350656	6	+	6	259	PF3D7_073130	
Pf3D7_14_v3	2207943	2208075	6	+	6	132	PF3D7_145370	
Pf3D7_14_v3	1717491	1717623	7	+	6	132	PF3D7_144240	
Pf3D7_05_v3	491925	492041	6	+	6	116	PF3D7_051150	
Pf3D7_07_v3	1350787	1351048	8	+	6	261	PF3D7_073130	
Pf3D7_14_v3	862709	862841	5	+	5	132	PF3D7_142070	
Pf3D7_13_v3	2338754	2338859	6	+	5	105	PF3D7_135900	
Pf3D7_14_v3	1446583	1447361	5	-	5	778	PF3D7_143570	
Pf3D7_11_v3	977274	977775	6	-	5	501	PF3D7_112480	
Pf3D7_02_v3	306675	306795	5	-	5	120	PF3D7_020760	
Pf3D7_07_v3	424191	424293	5	-	5	102	PF3D7_070940	
Pf3D7_08_v3	1032219	1032634	6	+	5	415	PF3D7_082330	
Pf3D7_14_v3	1022302	1022410	6	-	5	108	PF3D7_142620	
Pf3D7_10_v3	601196	601352	5	-	5	156	PF3D7_101490	
Pf3D7_10_v3	68337	68445	5	-	5	108	PF3D7_100110	
Pf3D7_14_v3	2526584	2526700	5	-	5	116	PF3D7_146230	
Pf3D7_14_v3	554485	555638	5	-	5	1153	PF3D7_141400	
Pf3D7_08_v3	670671	670780	5	+	5	109	PF3D7_081360	
Pf3D7_07_v3	545153	553405	5	-	5	8252	PF3D7_071230	
Pf3D7_14_v3	516906	517045	5	-	5	139	PF3D7_141280	
Pf3D7_11_v3	1955320	1955443	5	+	5	123	PF3D7_114900	

Supplementary Table 3.5 (Continued) | Putative *P. falciparum* circRNAs with at least five unique supporting reads.

Pf3D7_03_v3	430372	430551	5	+	5	179	PF3D7_031020	
Pf3D7_10_v3	892214	892325	5	-	5	111	PF3D7_102170	
Pf3D7_13_v3	2473986	2474136	7	+	5	150	PF3D7_136180	
Pf3D7_11_v3	465836	465980	5	+	5	144	PF3D7_111210	
Pf3D7_14_v3	2934564	2934698	5	-	5	134	PF3D7_147190	
Pf3D7_13_v3	366865	367075	5	-	5	210	PF3D7_130820	
Pf3D7_13_v3	151216	151714	5	-	5	498	PF3D7_130280	
Pf3D7_03_v3	724969	725857	5	+	5	888	PF3D7_031760	
Pf3D7_10_v3	1153629	1153730	5	-	5	101	PF3D7_102780	
Pf3D7_10_v3	1181901	1182006	6	+	5	105	PF3D7_102870	
Pf3D7_14_v3	1975985	1976111	7	-	5	126	PF3D7_144830	
Pf3D7_14_v3	862703	862841	5	+	5	138	PF3D7_142070	
Pf3D7_07_v3	755846	755954	6	+	5	108	PF3D7_071750	
Pf3D7_07_v3	1321921	1322044	5	+	5	123	PF3D7_073080	
Pf3D7_13_v3	112759	112877	5	-	5	118	PF3D7_130200	
Pf3D7_11_v3	390605	390925	5	-	5	320	PF3D7_110940	
Pf3D7_07_v3	276492	276609	5	-	5	117	PF3D7_070550	
Pf3D7_07_v3	1326576	1326678	5	+	5	102	PF3D7_073090	
Pf3D7_14_v3	978198	979167	5	+	5	969	PF3D7_142440	
Pf3D7_11_v3	226121	226532	6	+	5	411	PF3D7_110510	
Pf3D7_12_v3	117070	117230	5	-	5	160	PF3D7_120220	
Pf3D7_14_v3	289383	289558	5	-	5	175	PF3D7_140790	
Pf3D7_11_v3	1400498	1400624	5	-	5	126	PF3D7_113590	
Pf3D7_11_v3	914678	914807	6	+	5	129	PF3D7_112310	
Pf3D7_13_v3	2676871	2676991	5	+	5	120	PF3D7_136690	
Pf3D7_10_v3	436699	437481	5	+	5	782	PF3D7_101120	
Pf3D7_05_v3	492925	493049	5	-	5	124	PF3D7_051150	
Pf3D7_13_v3	2375163	2375284	5	+	5	121	PF3D7_135960	
Pf3D7_10_v3	423230	424021	5	+	5	791	PF3D7_101070	
Pf3D7_05_v3	313405	313576	7	-	5	171	PF3D7_050760	
Pf3D7_06_v3	600323	600470	5	-	5	147	PF3D7_061430	
Pf3D7_08_v3	687919	688033	5	+	5	114	PF3D7_081420	
Pf3D7_12_v3	162583	162709	5	-	5	126	PF3D7_120290	
Pf3D7_07_v3	396236	396338	5	+	5	102	PF3D7_070880	
Pf3D7_08_v3	1161061	1161578	5	-	5	517	PF3D7_082670	
Pf3D7_12_v3	63054	63235	5	-	5	181	PF3D7_120070	
Pf3D7_08_v3	572635	572752	5	-	5	117	PF3D7_081130	
Pf3D7_10_v3	162759	162930	5	-	5	171	NA	
Pf3D7_14_v3	142828	143029	5	-	5	201	PF3D7_140390	
Pf3D7_10_v3	457256	457421	6	-	5	165	PF3D7_101180	
Pf3D7_06_v3	111456	111609	5	-	5	153	PF3D7_060260	
Pf3D7_09_v3	1318502	1319012	5	-	5	510	PF3D7_093320	
Pf3D7_14_v3	127516	129502	5	+	5	1986	PF3D7_140340	
Pf3D7_12_v3	977371	977902	5	-	5	531	PF3D7_122410	
Pf3D7_14_v3	2207961	2208075	5	+	5	114	PF3D7_145370	
Pf3D7_10_v3	1424310	1424445	7	-	5	135	PF3D7_103590	

Supplementary Table 3.6| CircRNA GO term enrichment results

circRNA-specific GO term	circRNA-specific GO term p-value
cytosolic ribosome (sensu Eukaryota)	2.26E-06
cytosol	2.26E-06
cytosolic large ribosomal subunit (sensu Eukaryota)	5.11E-06
cytosolic part	5.15E-06
large ribosomal subunit	2.47E-05
ribosomal subunit	5.95E-05
ribosome	1.96E-04
structural constituent of ribosome	1.96E-04
macromolecule metabolic process	2.32E-04
macromolecule biosynthetic process	4.52E-04
protein metabolic process	1.27E-03
non-membrane bound organelle	1.32E-03
intracellular non-membrane bound organelle	0.00132
cellular protein metabolic process	0.00215
cellular macromolecule metabolic process	0.00228
ribonucleoprotein complex	0.0023
translation	0.00387
primary metabolic process	0.00413
gene expression	0.00552

Supplementary Table 3.7 | Divergent circRNA PCR primer pairs targeting nine tested loci. Yellow indicates successful experimental confirmation of predicted circRNA junction.

Target	cDNA1	cDNA2	Product (bp)	Divergent circRNA PCR primer sequences
Pf3D7_0909300_F	T20	Ring	161	TAA TAC GAC TCA CTA TAG GGtagtagtagtagAAGAAGAACAGGCTAGAAAG
Pf3D7_0909300_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagTTCAGGTGTCAATAAGGACT
Pf3D7_1474200_F	T32	Late Trophozoite	505	TAA TAC GAC TCA CTA TAG GGtagtagtagtagGTTATCCAACTACAGAGGAGA
Pf3D7_1474200_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagATATGAACCACTTGAGTATA TTCC
Pf3D7_1438400_F	T44	Schizont	324	TAA TAC GAC TCA CTA TAG GGtagtagtagtagGAAACATCAACTGCTAATATA TCTT
Pf3D7_1438400_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagCAAACAAACAAACACCATA AATAAA
Pf3D7_1026000_F	T20	Ring	591	TAA TAC GAC TCA CTA TAG GGtagtagtagtagAAGTTCAAGATGGAAGTGC
Pf3D7_1026000_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagATTCAAAATTATTATTTCAC AGAGGA
Pf3D7_0936800_F	TT8	Ring	571	TAA TAC GAC TCA CTA TAG GGtagtagtagtagAAAGGGTAATAATCAAGCGA A
Pf3D7_0936800_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagATTGTTATGTTCTTCTGGACT T
Pf3D7_1202600_F	TT8	Late Trophozoite	261	TAA TAC GAC TCA CTA TAG GGtagtagtagtagTGAAATGCCAAAGAAGAAAG A
Pf3D7_1202600_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagATCATCCTCTTGCTCCTTATA T
Pf3D7_0414900_F	T6	Ring	246	TAA TAC GAC TCA CTA TAG GGtagtagtagtagTTGAATCTGGTGCTTTGC
Pf3D7_0414900_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagATTGAGTTGCGGATAATGC
Pf3D7_0817700_F	TT8	Schizont	215	TAA TAC GAC TCA CTA TAG GGtagtagtagtagTGAACAACTTGAAATGGATTT ATT
Pf3D7_0817700_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagAAGGAACTTATATCTTGACT ATCTAG
Pf3D7_1314900_F	T20	Early Trophozoite	220	TAA TAC GAC TCA CTA TAG GGtagtagtagtagATGGAATTTGCTGGTAGAAA
Pf3D7_1314900_R				CGC CAG GGT TTT CCC AGT CAC GACtagtagtagtagTCTCTTACAACATCTTCAAC AA

References for Supplemental Material

1. Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG (2003) Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malaria Journal 2: 27.
2. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, et al. (2014) Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. BMC Genomics 15: 150.
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628.